Frameworx Best Practice

Big Data Analytics Guidebook

Big Data Analytics Solution Suite

GB979

Release 16.5.1

June 2017

Latest Update: Frameworx Release 16.5	TM Forum Approved
Version 5.0.2	IPR Mode: RAND

Notice

This document and the information contained herein is provided on an "AS IS" basis and TM FORUM DISCLAIMS ALL WARRANTIES, EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO ANY WARRANTY THAT THE USE OF THE INFORMATION HEREIN WILL NOT INFRINGE ANY OWNERSHIP RIGHTS OR ANY IMPLIED WARRANTIES OF MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE.

TM FORUM invites any TM FORUM Member or any other party that believes it has patent claims that would necessarily be infringed by implementations of this TM Forum Standards Final Deliverable, to notify the TM FORUM Team Administrator and provide an indication of its willingness to grant patent licenses to such patent claims in a manner consistent with the IPR Mode of the TM FORUM Collaboration Project Team that produced this deliverable.

The TM FORUM invites any party to contact the TM FORUM Team Administrator if it is aware of a claim of ownership of any patent claims that would necessarily be infringed by implementations of this TM FORUM Standards Final Deliverable by a patent holder that is not willing to provide a license to such patent claims in a manner consistent with the IPR Mode of the TM FORUM Collaboration Project Team that produced this TM FORUM Standards Final Deliverable. TM FORUM may include such claims on its website, but disclaims any obligation to do so.

TM FORUM takes no position regarding the validity or scope of any intellectual property or other rights that might be claimed to pertain to the implementation or use of the technology described in this TM FORUM Standards Final Deliverable or the extent to which any license under such rights might or might not be available; neither does it represent that it has made any effort to identify any such rights. Information on TM FORUM's procedures with respect to rights in any document or deliverable produced by a TM FORUM Collaboration Project Team can be found on the TM FORUM website. Copies of claims of rights made available for publication and any assurances of licenses to be made available, or the result of an attempt made to obtain a general license or permission for the use of such proprietary rights by implementers or users of this TM FORUM Standards Final Deliverable, can be obtained from the TM FORUM Team Administrator. TM FORUM makes no representation that any information or list of intellectual property rights will at any time be complete, or that any claims in such list are, in fact, Essential Claims.

Direct inquiries to the TM Forum office:

240 Headquarters Plaza,
East Tower – 10th Floor,
Morristown, NJ 07960 USA
Tel No. +1 973 944 5100
Fax No. +1 973 944 5110
TM Forum Web Page: www.tmforum.org

Table of Contents

List of Figures

Executive Summary

Big Data Analytics is having a huge transformative impact on many industries. Combining more effective lower cost data storage and presentation with better data processing mechanisms has allowed many industries to unlock additional value from their existing data. This is especially the case for digital service providers where there is an abundance of available data as well as the desire for such service providers to investigate new revenue sources.

TM Forum, in its role in steering technology, terminology, standards and best practices within the digital services space, has an active Big Data Analytics Project group to lead the introduction and adoption of Big Data Analytics within this space. The group initially released their first document, TR202 to provide a standard architecture reference model for Big Data Analytics. This document introduced the concept of the reference implementations and also outlined an initial set of business-level use-cases to help with the deployment of Big Data Analytics in Telecoms projects.

This standard, TM Forum Big Data Analytics Guidebook (GB979) along with its Addenda: Addendum A - Big Data Analytics Use Cases (GB979A); Addendum B - Big Data Analytics Building Blocks (GB979B); Addendum C – Privacy Risk Scoring (GB979C); and Addendum D - Analytics Big Data Repository (GB979D) are intended to follow-on from and supersede TR202 by providing many additional Big Data reference use-cases as well as describing how Big Data Analytics can help unlock business value in the digital services domain.

The Big Data Analytics Guidebook furthermore describes a clear path to help implement Big Data Analytics projects. This path, named "Big Data Analytics Business Value Roadmap" consists of the following steps:

1. Service Providers select the business cases that they wish to use Big Data Analytics mechanisms to implement - from the 50+ that are described in GB979A.
2. Service Providers will identify functional modules within these use-cases that need to be implemented. These re-useable functions are called "Big Data Analytics Building Blocks" (ABBs). ABB is a new concept introduced within this document with the goal to provide references to specific Big Data Analytics implementations. The appendix GB979B provides an initial catalog of many of these that can help companies accelerate the delivery of use-cases as well as increase ROI and lower risks.
3. Finally by providing a map of technologies and data sources along with Building Blocks to use-cases this document provides a reference model to help describe an end-to-end framework for realizing business value for service providers from Big Data Analytics.

Following on from this we hope that this guidebook will be improved over time to describe addition use-cases, Building blocks and refine the Reference model and Framework.

It is the hope of the Big Data Analytics Project group that this guidebook will help CEPs to answer both the "why" in Big Data Analytics for communications but also start to answer the "how" value can be realized using these techniques.

Document Structure

Executive Summary: Summarizes the main points from the document and highlight the problem statement being addressed, the main results, the conclusions drawn and the next steps as appropriate.

Introduction: overview of this the document and outlines its structure and defines essential terms used in the document.

BDA Business Value Roadmap: best practice on how to leverage big data analytics to realize business values through a six-step process

Big Data and Big Data Analytics: definitions of Big Data and high level description of Big Data Analytics techniques

Analytics Big Data Repository: introduction to ABDR construct; further details in GB979 Addendum C

BDA Reference Model: functional components of a Big Data Analytics platform within a Big Data Solution ecosystem

Big Data Repository Vertical: description of the layer within the reference model which provides storage of all data within the big data platform

Data Governance Vertical: description of the layer within the reference model which provides privacy, security and compliance functions within the big data platform

Data Flow: overview of the different types of data flows typically encountered within a Big Data Solution ecosystem

Additional Details on Synthetic Data Model: discussion of a fictitious set of data having the same characteristics as a set of true data, in such a way that it is not possible to re-identify the initial true users

Administrative Appendix provides document revision history, acknowledgements for work completed and information about the TM Forum.

1. Introduction to Big Data Analytics Guidebook

1.1. Foreword

From TM Forum publication: ***BIG DATA ANALYTICS: EXTRACT VALUE FROM THE DATA TORRENT***

Big data remains a hugely important topic this year across virtually all industries and the communications sector is no exception. Business leaders, especially in marketing and IT management, are bombarded with publicity about the crucial role of importance of big data, and they are very optimistic about the possibilities buried in the zettabytes (1 zettabyte equals 1 trillion gigabytes) of data stored today. Indeed IDC estimates that there are 4.4 zettabytes in captivity today, which is expects to expand ten-fold to 44 zettabytes by 2020. Of course, for all the potential, organizations are struggling to find a successful approach, daunted by the ever-increasing torrent of big data and the predicted failure of any organization foolish enough to ignore the opportunity. In fact, for many the collection of technologies that comprise big data are stalled in the Trough of Disillusionment in Gartner's 2014 hype cycle, or sliding down into it. Big data has all the characteristics of an emerging technological phenomenon: the scope is huge; the definitions hazy and sometimes conflicting; and the complexity is unprecedented.

1.1.1. Making progress

While there are valid arguments from optimists and pessimists alike, many service providers are making progress in what is still an early-stage market. Accordingly, we believe that big data justifiably remains among the most pivotal, forward-looking topics for business and IT organizations within all kinds of service providers today, and it will remain among their top issues and initiatives for some time. This importance is not only reflected in the potential value to be derived from the huge amount of data they generate and are exposed to, but in the criticality of getting information and decision management right if they are to transform to become digital, customer-centric businesses. The key to success, of course, is to unlock the value in that data.

1.1.2. Defining big data

There is no generally accepted formal definition for big data – and the myriad 'definitions' that exist are typically characterizations. Where does that leave us? Well, supporting our pragmatic approach, we believe there is no 'one size fits all' definition. In fact the most useful definition for big data analytics, in our view, would be one that supports the identification, design and deployment of strategies, processes, skills, solutions, tools and data which can provide actionable intelligence to deliver business value.

1.2. Introduction

The focus of the TM Forum's Data Analytics guidebook is to provide TM Forum members with approaches, tools, and most importantly, a common language to accelerate, streamline and de-risk data analytics projects.

Service providers today gather extensive data from multiple sources – credit scores, call logs, customer support call transcripts, social media activity, browsing history and service history – just to name a few. This data carries the potential for significant value that can enhance the way service providers do business. But the value needs to be unlocked and unlocking that value in a systematic fashion will save time and money.

The best practices in this guidebook family provide the industry with this systematic approach and in turn deliver a firm foundation with which to leverage data analytics solutions to keep improving the way service providers and their suppliers do business. They aim to deliver well thought out approaches and guidance that helps service providers define direction and unlock the value from the data they collect. They do that through providing a common language to be used internally in service providers, but also to help suppliers and integrators have a common language for data analytics with their customers. Having a common language means that service provider projects can run more smoothly internally, suppliers will have a clearer understanding what SPs are asking for, and suppliers can drive the discussion about priorities with their customers using industry agreed upon terminologies.

The guidebooks take a pragmatic and practical approach and are designed to give any service provider or supplier a running start in their big data analytics programs. In this volume which is the overarching best practice, the content consists of:

- A business value roadmap to help you plan your data analytics project step by step, leveraging TM Forum's data analytics best practices.
- A reference model to define the functionality needed to implement a data analytics solution. At the bottom of the model are the typical data sources and at the top are the value added uses for the data. In the middle are all of the functions that are needed to extract the value out of the data.
- An introduction to the library of use cases that identify over 60 ways that service providers can extract value from their data. Each use case has a business canvas, a set of metrics associated with it, mappings to the TM Forum Business Process Framework, mapping to the TM Forum Customer Experience Lifecycle Model, and a list of data sources. We have use cases related to personalized services, product performance, proactive care, churn, network management, and much more. Again, this is a common language for discussing priorities for data analytics implementation, and a way to accelerate a project.

Everything in this overview document and the associated appendices was created collaboratively by TM Forum members based on their real world experiences. Contributions to this work are helping the whole industry move forward and we hope you will find benefit in the new language you will be learning.

2. Big Data Analytics Business Value Roadmap

2.1. Big Data Analytics Business Value roadmap

Implementation of big data analytics is linked to the digital transformation of an enterprise. Many factors must be considered when charting a course to take full advantage of big data analytics, including:

- selection of high value use cases for implementation,
- deliberation of strategies such as data repository and other IT approaches,
- identification of a recognized champion with ability to define and implement strategies,
- data governance and privacy concerns.

TM Forum defines a Big Data Analytics (BDA) Business Value Roadmap as an industry best practice on how to leverage big data analytics to realize business values through a six-step process *illustrated below in Figure 1*. TM Forum recommends this process for all organization with plans to leverage big data analytics to unlock new and undiscovered business values. It is important for enterprises to realize that with a few exceptions (such as some truly distributed, scalable machine-learning analytics algorithms), the technologies exist today to solve their big data challenges. The real difficulty lies in identifying the right BDA approaches and technologies which best fit the needs of high value business use cases as measured by ROI, NPS or other metrics. This is where BDA Business Value Roadmap provides the most value. If this is your first time reading this document, read through this section first to understand the approach that is being mapped out, but do not dwell on the details of any given element. Then read through the rest of the document to get more in depth on each of the steps and resource, and finally come back and re-read this section to put all the pieces together.

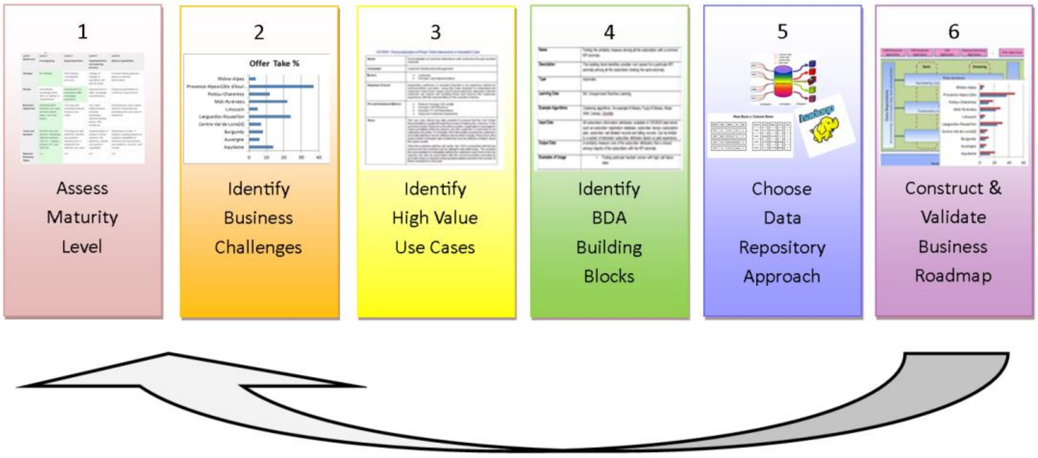

1
Assess Maturity Level
2
Identify Business Challenges
3
Identify High Value Use Cases
4
Identify BDA Building Blocks
5
Choose Data Repository Approach
6
Construct & Validate Business Roadmap

Figure 1: Business Value Roadmap

2.1.1. Step 1: Understand Your Big Data Analytics Maturity Level

Goals for an enterprise embarking on a new or additional data analytics program are likely to focus on:

- obtaining optimal business value
- planning and executing a program of continuous improvement around their big data analytics strategy.

In order to accomplish these goals, it is useful to have a means to assess the organizations' maturity across key defining factors of big data analytics implementation requirements. The TM Forum has created a model for this assessment which consists of maturity levels across each of 4 Big Data Analytics Dimensions. The model is based on the Customer Lifecycle and the Frameworx business model, providing a structured and practical approach to creating this best practice.

It should be noted that an enterprise is seldom monolithic in its level of maturity across all dimensions or attributes. The path to obtaining business value and realizing comprehensive improvement in implementing a big data analytics strategy may vary according to factors unique to the circumstances and environment in which the enterprise finds itself at any given time.

Big Data Analytics Maturity Level Matrix

Level / Dimension	**Level -1** **Ad hoc**	**Level -2** **Tactical**	**Level -3** **Competitive**	**Level -4** **Differentiating**	**Level -5** **Innovative Leader**
Definition	Rudimentary, non-standard execution, designed to meet immediate needs only Siloed groups or departments of an organization leverage elementary analytics for gathering unsystematic data insights	Standardized departmental focus with a primary objective of realizing functional or operational efficiency Data analytics leverages intra-departmental IT and operational synergies, through evolving processes, tools and standard guidelines	Standardized organization-wide inter-departmental focus. Primary objective is to leverage analytical insights for making informed decisions around attaining service optimization and operational efficiency Data analytics is institutionalized as a tool to remain competitive and enable business strategy	Mature data analytics insights, enabling initial levels of service differentiation and consistency with industry best practices and standards Growing capability to deliver personalized, contextualized and customer-oriented insights to enable better end-customer experience, operational excellence, profitability and revenue streams	Mature and innovative data analytics insights enabling consistent service differentiation, new offerings and innovative customer experience Establishing next-generation best practices through innovative data analytics implementation Data analytics enables innovative, new opportunities and avenues for best-in-class customer experience, new revenue streams and continuous operational

Level / Dimension	Level -1 Ad hoc	Level -2 Tactical	Level -3 Competitive	Level -4 Differentiating	Level -5 Innovative Leader
					excellence
Business strategy	Data Analytics is not part of organization's business strategy	Data analytics realization is largely used to solve tactical business intelligence issues and is not aligned to organizational business strategy	Data analytics is an integral part of organization's business strategy Organization wide acceptance to leverage insights from data analytics	Data analytics plays an influential role in the organization's business strategy Data analytics insights enable validation and justification of identified strategic priorities and occasional course corrections	Data analytics enables continuous business model innovations and formulation of transformational business strategy Data analytics insights enable informed customer centric business, operational and IT initiatives, and ROI guidance
Business & Functional Capabilities	Elementary analytics confined to "as-is" analysis of historical structured data per ad-hoc requirements	Mostly descriptive and informative in nature with no focus on predicting future outcomes Use of data analytics extends only to financial, operational and regulatory reporting	Mature descriptive and informative capabilities with increasing focus on business, functional and operational predictions and customer experience management. Data analytics enables data driven decision making with decreasing dependency on human inputs Greater emphasis on predictive analysis	Maturing predictive capabilities with increasing focus on prescriptive remedial actions Data analytics insights further reduce the human component for deducing business forecasts and providing prescriptive remediation Higher degree of automation and prediction	Mature prescriptive capabilities with increasing accuracy and consistently out-performing human remediation and forecasting Innovative data analytics enable new opportunities and avenues for best-in-class customer experience, new revenue streams and continuous operational excellence
Technology Platform and Systems	Data analytics platform not present Low functionality, tools and	No coherent data analytics platform strategy in place. Multiple departments may have individual	Data analytics platform, systems and tools are prescribed in a well-defined enterprise	Sophisticated data analytics platform, systems and tools on a well-defined reference enterprise	Highly resilient platform, systems and tools support seamless integration of new data sources and

Level / Dimension	Level -1 Ad hoc	Level -2 Tactical	Level -3 Competitive	Level -4 Differentiating	Level -5 Innovative Leader
	systems leverage historical BI for basic reporting using graphs, what-if analysis, t-test, f-test, etc. Data extraction based on ad-hoc needs	platform approaches and selection of systems and tools to meet tactical needs. Basic analytics algorithms do not support predictive analysis	architecture. Analytics tools and algorithms enable predictive analysis. Awareness and operationalization of appropriate platform, vendors and analytics infrastructure to implement the selected use cases	architecture to effectively deal with challenges of data volume, veracity, variety and velocity. Tools and algorithms support prescriptive analysis.	cutting edge algorithms catering to future and next generation customer and enterprise demand.
Architecture	There is no single coherent data analytics architecture. Only low volume of structured data can be managed.	Low, but improving, level of architectural maturity; can manage only low & medium volume of structured data	Dedicated effort to define an effective information architecture able to manage a high volume of semi-structured and un-structured data types	Well rationalized architecture is defined in accordance with enterprise strategy and ability to deal with volume, velocity, variety and veracity requirements.	Leading edge and continuously improving enterprise architecture capabilities and analytics platform able to adapt to changing business needs.
Data Management	There are no defined data management policies and guidelines Data is managed on an as-needed basis	Basic data management policies and guidelines in place at the department level Addresses low volume structured data only	Organization wide standard data management practices and guidelines to leverage inter-departmental synergies Focused on leveraging structured, semi-structured and un-structured data from organization wide data sources	Mature data management practices and guidelines are well defined Seamless integration of various internal and external data sources, across data analytics lifecycle	Innovative data management practices are well integrated with business strategy and changing customer demands for delivering seamless customer experience Data becomes new line of business with opportunities for new revenue sources from 3rd parties
Processes	No data analytics processes	Basic process flows and models for	Standard inter-departmental data analytics	Standard and end to end automated	Robust and automated exception

Level / Dimension	Level -1 Ad hoc	Level -2 Tactical	Level -3 Competitive	Level -4 Differentiating	Level -5 Innovative Leader
	exist	departmental data analysis and reporting are in place	processes are in place Preparation for automation and data analytics optimization. Some component based automation is in place	data analytics processes are in place across the organization	handling Continuous evolution to meet current industry best practices for optimum benefit realization
Budget & Financial Tracking	No dedicated budget and financial tracking	Data analytics is a departmental priority and is funded at department level Financial returns and ROI are not tracked centrally	Budget allocations for procurement and implementation of data analytics solution are made centrally Initial criteria for data analytics financial returns are defined and monitored	Budget allocation is made with organization wide perspective and in a cross-functional manner Detailed business case driven criteria including benefit tracking and ROI are in place	Budget allocation is continuously re-evaluated using metrics to optimize the data analytics investment
Governance	No data analytics governance defined	Data analytics governance is defined at the department level Governance includes definitions of data sources (inbound and outbound), task ownership and managing the data effectiveness	Data analytics governance is defined at organization level and driven by centralized data analytics council Standard governance policies, procedures and guidelines are defined for driving seamless inter-departmental synergies	Data analytics governance best practices are used for cross functional projects with increasing consistency Automated tracking and monitoring of privacy, security and compliance for effective data usage	Data analytics governance is tightly integrated with all aspects of business strategy and operations Data analytics governance is consistently used organization wide across all projects
Organizational Readiness & Management	Leveraging and implementing analytics is mostly a choice of an individual or	Awareness of the importance of standard analytics exists at the department level, however	Organization wide acceptance to utilize data analytics insights for delivering against wider business	Data driven decision making is routinely operationalized at organization	Metrics-based refinement of policies where results are measured around meeting customer

Level / Dimension	Level -1 Ad hoc	Level -2 Tactical	Level -3 Competitive	Level -4 Differentiating	Level -5 Innovative Leader
	a department, and has little effect on the way the broader organization operates Coincidental knowledge; no organized training provided	in general the company, at a cultural level, is largely unaware of analytics ROI Starting to develop awareness, skills and knowledge training for small analytics assignments	objectives Dedicated cross-functional analytics resources with well-defined competency catalog and future skill/re-skill training roadmap	level Data analytics activities influence employee motivation for personal and organizational development	experience and other strategic goals through leveraging data analytics insights Organization's resources continually optimize, innovate and break-away from non-data driven modes of business operations
KPIs/Business Metrics	No KPIs/KQIs or metrics identified for use in measuring product or service performance or quality. No metrics identified for measuring effectiveness of data analytics utilization	Initial KPIs/KQIs (mostly operational) identified for siloed performance and quality measurements. Customer experience not readily tied back to KPIs and KQIs. Metrics are identified (mostly operational) to understand benefits of using data analytics	Cross functional KPIs/KQIs are identified and cataloged at organization level. KPIs/KQIs used to measure QOE as an indicator of customer experience and drive business decisions at an organizational level Cross functional metrics are used to measure data analytics effectiveness.	KPIs/KQIs/QOE & metrics are benchmarked to industry standards and used to drive business decisions at an enterprise level. Cross functional metrics are used to measure and continuously improve data analytics effectiveness.	KPIs & metrics are further optimized for better business enablement and innovative customer experience. New services and areas of innovative experience are automatically raised through metrics.

The following attachment is an automated toolkit which enables companies to assess their maturity on all of these dimensions. It is an automated toolkit which can give quick insights into where gaps exist and what specific areas of improvement are most crucial for an organization.

BDA_Maturity
Assessment Question

2.1.2. Step 2: Identify Business Challenges which can be Improved with Big Data Analytics

Understand your enterprise's business challenges that can be improved by using big data analytics. The key here is to ensure that data analytics projects are targeting high priority issues and needs in the organization and are not just analytics for analytics sake. These challenges can be canvased from multiple organization within the enterprise, e.g. Marketing, Product Management, Operations, Suppliers and Partners. The challenges can then be cross-referenced to the TM Forum Business Process Framework (eTOM) in order to define them in a common fashion. Some examples of areas where challenges that can be solved with data analytics are listed below.

- **Strategy**
  - Provide real-time personalized offers
  - Entice additional usage
- **Product Portfolio**
  - Optimize product performance
  - Analyze product introductions
- **Operations**
  - Provide proactive care
  - Drive network repair based on customer experience
- **Profitability**
  - Predict churn propensity
  - Assure revenue
  - Detect fraudulent usage

2.1.3. Step 3: Identify High Value Use Cases and Refine

In this step, applicable use cases are selected to address challenges identified in Step 2; e.g. improving customer experience, developing new revenue opportunities, enhancing network operational efficiency, to name a few. Typically, organizations may already have several business use cases identified at this stage. They can further consult the TM Forum Big Data Analytics Use Cases documented in GB979A for substantial additional resources. Of course, existing use cases should be refined, if necessary, to reflect the particular business challenge, organizational environment, data availability and any other factors.

Whether starting from scratch or refining an existing use case, TM Forum welcomes your contributions to the Use Case portfolio in GB79A.

Once use cases are identified and refined, they can be prioritized according to their business value using the Ostervander canvas found in each TM Forum use case or using the Osterwalder Canvas as a template for a new use case.

A use case template document is used to collect all data about a use case. All existing use cases have this template completed. The use case is described in detail including its purpose and business value. Part of this description can be the relation of the use case to other models; for example, eTOM.

Name: The name of the use case
Horizontal(s) or Vertical(s): The horizontal or vertical areas of the TM Forum Business Process Framework process areas that the use case touches
Actors: Entities involved in the use case, e.g. “Customer”, “CSR” etc.
Business Drivers: A short description of the solution that describes its core value and why a service provider would be motivated to implement this use case;
Business Metrics: A list of the business metrics that this use case impacts. The metrics used to describe each use case are those defined in the TM Forum Business Metrics Specification GB935A
CxLC Stage: The stage of the customer experience lifecycle (CxLC) that this use case impacts
Customer Experience Metrics: A list of the customer experience metrics that this use case impacts. The metrics used to describe each use case are those defined in the TM Forum Customer Experience Management Lifecycle Metrics specification GB962A
Story: A description of the flow of the use case

2.1.4. Step 4: Identify BDA Building Blocks

To realize use cases chosen in step 3, the service provider should know how to analyze each use case for things such as input data, output data, how to process the underlying data, and which algorithm to analyze the data with. Some use cases have duplicate or similar input and output data, as well as a similar algorithm to process data, so they can use the same functional modules.

GB979B identifies a suite of building blocks that are needed to realize the use cases identified in Step 3. Each of these BDA building blocks performs a fundamental function that is use case independent to maximize reuse.

Service provider can choose the building blocks documented in GB979B, if no such building blocks have been documented, please consider documenting them and contributing them back to the GB979B standard.

Each of the big data analytics building blocks are documented in a structured manner using the following attributes:

- **Name:** The name of the building block;
- **Description:** A high level description of the building blocks function;
- **Type:** One or more of **Automatically Learned**, **Manual Specified**, or **Programmatically Described;**
- **Underlying Data:** In the case of models that are Automatically Learned this attribute describes the data used to build the model that underlies the building block. In the case of Manually Specified or Programmatically Described this attributes describes any data used to power the building block;
- **Example Algorithms:** Documents the type and provides examples of the sorts of machine learning algorithms that can be applied on the “input data for learning” in order to produce the core model that underpins the building block;
- **Input Data:** Describes the input data provided to the building block at runtime;
- **Output Data:** Describes the output produced by the building block at runtime;
- **Example of Usage:** Gives one or more examples of real usage of this analysis model in order to deliver value to the CSP.
- **Related Use Cases:** Example use cases from GB979A related to this building block

Step 5: Choose the Right Big Data Repository Approach

Among the big data analytics use cases identified, the service provider should be able to determine which infrastructures to put in place to store and handle big data and which underlying technology to be used to access and process these data.

- The Analytics Big Data Repository (ABDR) provides definition of the entities
- The current set of big data analytics use cases provide list of mandatory and optional data sources and the logic of the requirement.
- Based on such specifications the service provider will be able to determine the data sources and entities to be used, taking into account that reuse of entities from big data repositories across the various infrastructures is absolutely a mandatory approach. That means that with multiple use cases requiring the same data with the same attributes as defined above, the data should be used in a unified manner.

The following attributes are very important and should be defined for each data entity in the ABDR to help determine how the data will be used by each use case and which infrastructure to use:

- Data availability - define how the data will be available: via streaming (when it is expected to access this data in real-time), or via batch (when it is expected to access this data in near real-time, with an accepted lag time)
- Data expandability - ensure there is the capability to keep adding solutions that use this data
- Data resolution - in how granular or aggregated of a fashion will the data will be retained? For drill down/up/through purpose identify if the data is hierarchical.
- Data volume - how many rows of data will there be? This will depend on the data itself (usages, events, CDRs are typical data with very high volume over time)
- Data retention - how long will the data be retained in the ABDR for the various purposes (for some use cases it would be needed forever, for some other only for 1 year). Define for each type of data how long it will be kept for each level of granularity.
- Data quality - how clean is the data? Some use cases need to analyze cured data to consider only correct events be able to generate expected KPIs, while other use cases need to access data with errors to determine KPIs on generation of errors and how the production systems led to such a situation.

Depending on these data attributes, the service provider will also be able to define the way they will access the data from the data sources. It could be through ETL/ELT processes but also through streaming. The definition of data flows will be extremely important in order to describe how the data are flowing from ecosystem data sources to the various analytics infrastructure components.

The use cases identified by the service provider will also guide the selection of the infrastructure through the application processing needs. For example Policy Analytics will mainly require data analysis on large number of events, CDRs and usage to discover new trends and potential opportunities to develop new policy plans or new products. Such cases might require ad-hoc reporting/analysis and drill down/up/through, and exploitation of the data through advanced data visualization capabilities. Another example would be to determine customer behavior over time regarding top-up and data consumption to identify what is the best offer to suggest to the most valuable customers (highest uplift long term revenue). For that purpose, the solution would need to handle time series analysis, churn prediction, long term revenue calculation, and determination of the best time to send a marketing message for a maximal ROI. In this case a Hadoop solution could provide a better solution for data scientists to address the expected results.

Once defined, each ABDR entity must be categorized by type of infrastructure through which it should be available:

- Hadoop storage
- Columnar DB
- Real-time access to data through streaming and complex event processing

Use cases can require the use of multiple data based on multiple analytic infrastructures to deliver the expected analytical processing.

Once all the Analytics Big Data entities' attributes are defined, the service provider will be able to define the Analytics Big Data Repository approach to handle their analytics use cases.

2.1.5. Step 6: Construct and Validate BDA Business Value Roadmap

At this point it is necessary to pull together all technologies and information resources identified by the Use Cases and Building Blocks. The components should be mapped to the BDA Reference Model and validate that all necessary areas have been addressed.

Combine all artifacts together for the use case using the steps above. This is the first pass BDA Business Value Roadmap. It may be prudent to execute a Proof of Concept. Be sure to go back and validate the result with the use case. There may be a few more iterations to go through so that the Business Value Roadmap can be finalized.

The final step in the overall process is to reassess the business situation after the BDA project has been implemented and to decide on further activities. For example, the results might lead to a decision to loop back and select further actions or refinements. This should also include evaluation against the business priorities in the business canvas and business metrics in the use cases; i.e. determination of ROI, improved NPS or other metrics for the result's success.

The final part of the use case description is a collection of implementation stories. These are documented experiences of organizations that have applied the recommended solutions. They can describe how successful they were, what obstacles did appear, how they were solved and where the organization did deviate in a certain way from the proposal. Using this approach the organization can more rapidly build on past experiences when building new solutions.

2.1.6. End-to-end Business Value Roadmap Example using Use Case: MS-SAM-3: Real-time Personalized Offers Based on Location

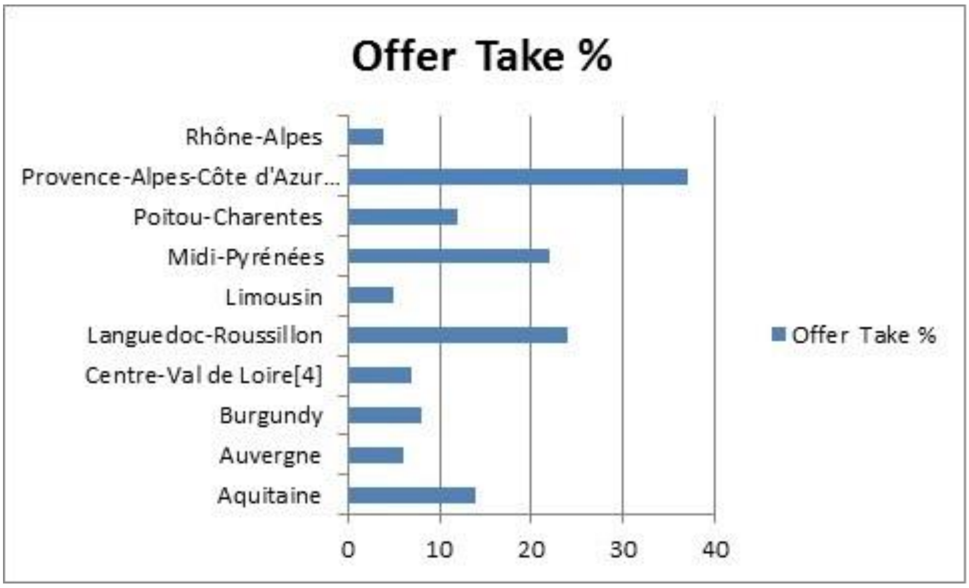
The following example walks through the 6 step end-to-end process described above.

Example Step 1: Understand Your Big Data Analytics Maturity Level

XYZ Ltd. is a multinational conglomerate providing various digital services to its customers. It has undertaken several Data Analytics projects in different departments within its enterprise, but has not yet developed a cohesive big data analytics strategy, business objectives or tools. Several key personnel have been developing the knowledge and skills needed to undertake the challenge. A corporate champion for big data analytics has decided to use the TM Forum process as a guide to implementing their approach. As a first step, she makes an honest assessment of XYZ's maturity level in these key areas.

Example Step 2: Identify Business Challenges which can be Improved with Big Data Analytics

Many XYZ business units were canvased for their business challenges. One that stood out came from the Marketing and Offer Management teams. They have noticed that service offers are often successful in some geographic areas, but not others. This business challenge is selected as a test case to determine if they can increase offer marketing effectiveness using big data analytics.



Example Step 3: Identify High Value Use Cases and Refine

Based on the business challenge, the following use case, S-MOM-T4, was found in TM Forum's Big Data Analytics Guidebook Use Cases document (GB979A) and determined to be a good fit as is, despite the fact that it was originally intended to provide a much finer grained mobile location than the initial pilot project would be using.

Name:	Personalized Marketing to Mobile Subscribers Based on Customer Location
Horizontal:	Marketing and Offer Management
Actors:	• Customer

Business Drivers:	Mobile Marketing is both profitable and risky. When relevant marketing messages are pushed to customers, they are useful information that enhances customer experience; on the other hand, when they are not relevant to customers, they become spam and customers are at risk of churning. Leveraging Big Data Analytics, Mobile Marketing can be triggered by customer location changes and thus increase the chance of relevancy to what customers need. At the same time, it becomes a differentiator for communication providers compared to other marketing campaign services.
PI's and Business Metrics:	• Increased Revenue • Increased Offer Acceptance Rate • Improved Customer Experience
Story:	This use case utilizes big data analytics to ensure that the mobile marketing

	messages, from a catalogue of most relevant, pre-defined campaigns, are sent only when customer arrives or is about to arrive at certain pre-defined geo-fenced locations. Due to real-time nature of the marketing offers, the Campaign Management System is required to send out the messages to customers within a few minutes of before or after customer arrival at the location. In case of predicting customer locations, their information such as customer demographics, web browsing history, call history and social media records can be used to augment the analytics to make the location predictions more accurate. Over time, the system can build a profile of the customer locations and distinguish different locations with different labels so that the right advertising is sent to customer at the right location.
Required Data Sources:	• Offer Catalog • Availability & Eligibility Rules • Mobile Location Information • Customer List (opt-in or opt-out)
Optional Data Sources:	• Call Detailed Records • Social Media Records • Web Browsing History

Example Step 4: Identify BDA Building Blocks

The following BDA Building Blocks have been identified and documented in Big Data Analytics Guidebook Building Blocks (GB979B) with predictive analytics algorithms.

- CL3: Customer Location Prediction
- CL4: Customer Key Location Profiling

CL3: Customer Location Prediction

Name:	Customer Location Prediction
Description:	This building block is used to predict where a customer is going to be at a specific time in the future or predict when a customer is likely to be next in a particular location
Type:	Automatically Learned
Underlying Data:	A set of coarse grained customer location information; relationship between customers (call logs, social media connections etc.) for advanced modeling
Example Algorithms:	Supervised learning methods using Spatiotemporal contexts of customer historical locations[1] or more advanced modeling taking into account of the result of social analysis[2]
Input Data:	Customer’s current location and/or time
Output Data:	Customer’s next location
Examples of Usage:	• Predicting when a customer will be at a specified location in order to send a relevant targeted offer before they arrive • Predicting where a customer is likely to be at a specified time in order to send a relevant targeted offer before they arrive
Related Use Cases	• S-MOM-T4 • Emergency messaging/crowd control

	• Traffic Alert
Implementation Guide	**Data Source** The available data sources are • Cell Tower attached to • Database of cell tower locations and directional information • Wi-Fi Base Station attached to • Database of Wi-Fi base station locations • GPS data • Radio signal strength and timing **Data Ingestion** There are several levels of location information with increasing accuracy. • Single Cell • Multiple Cells • Wi-Fi • GPS The level of accuracy required is highly dependent on use cases. While cell level location accuracy is acceptable for sending location-based advertising for the nearby coffee shops, it becomes intolerable for pinpointing customer locations for traffic directions. **Data Management** Location data volume can be large based on the number of customers tracked and how long the records are kept for analysis. Data Management Layer can compress from stream of locations and generate less than 100 location records a day with non-duplicate location and time when the location is reported. This can still add up to billions of records per day for a large CSP. If the analysis requires beyond the spatiotemporal information such as social network information below, the storage will quickly expand into Big Data realm and require non-traditional databases to host the information. **Data Analysis** For location prediction, data is accessed in two phases. *Data Modeling Phase* During the Modeling Phase, data is accessed in batch mode with location, time, and other related information retrieved and processed for all customers. This is done typically on a massive parallel processing platform (e.g. Hadoop Cluster with MapReduce). The resulting model contains the probability of next location per current location for every customer. It then can be used in real-time to predict the customer's next location as customer's location changes. Most of the modeling algorithms use supervised learning based on customer location first. Then they add on additional information such as time to increase prediction accuracy. Reference[1] describes an example algorithm that takes into spatiotemporal context of customers at the same time. The algorithm considers both prior location and prior time period (after making some assumptions) when calculating the combined probability on the customer's next location. Social network information can also be added to improve location prediction accuracy. In this case during modeling, training data includes the location

	and time of customers' "friends". Reference[2] algorithm's prediction accuracy is 1-2 orders of magnitude better than the same algorithm without taking social network information into consideration. Higher prediction accuracy usually means more costly to implement in Big Data Analytics environment with larger training sets and expensive processing. In the end, the added cost will have to be justified by the use cases that include this building block. *Location Prediction Phase* As new location information streams in for each customer, the model is applied to each new updates to predict the next likely location for the customer. Periodically, the information can be stored off to a Big Data repository for batch processing.

CL4: Key Location Profiling

Name:	Key Location Profiling
Description:	This building block utilizes customer location information that can be gathered over their lifetime in order to learn the key location information for this customer.
Type:	Automatically Learned
Underlying Data:	Potentially a set of labels that the CSP would like to identify locations for and a set of coarse grained historical customer location information
Example Algorithms:	Clustering algorithms, for example K-Means, based on Spatial distances and time of the day
Input Data:	Customer's current location and time
Output Data:	Location labels
Examples of Usage:	• Ensuring that targeted offers are sent to the customer when they are at home or socializing and not bothering them while they are at work. • Ensuring that outbound interactions with the customer are made when they are in the right location, at work for business customers or at home for individuals.
Related Use Cases	• S-MOM-T4 • Location-Based Fraud Detection • Social Networking Apps
Implementation Guide	**Data Source** Same as CL3 **Data Ingestion** Same as CL3 **Data Management** Same as CL3 **Data Analysis** For location profiling, data is accessed in batch mode with location and time for all customers. This is done typically on a massive parallel processing platform (e.g. Hadoop Cluster with MapReduce). The resulting model contains the key locations with classification labels (e.g. home or office) for

	every customer. It then can be used in real-time to provide location label for each key location when location updates for the customers come in. Typical modeling algorithms use unsupervised learning, such as different clustering algorithms, based on customer locations and time they spend at each location. The longer customer spends time in one location, the more weight it is given in the clustering for identifying the clusters. After location clusters are identified, the time range customer spends in each cluster in considered for labels. For example, the location that the customer spends the most time between the hours of 10pm and 6am may be their home, the location where they spend their most time from 9am to 5pm may be their work, and the location where they spend 8pm to 11pm on the weekend may be where they socialize with friends. This building block may produce the customer's top locations, or may label the locations as needed by the CSP. In order to process months of location data at billions of records per day efficiently, clustering algorithms optimized for Big Data Analytics like canopy clustering are used. As new location information streams in for each customer, the model is applied to each new updates to label the new locations if they match a labeled location.

Example Step 5: Choose the Right Big Data Repository Approach

Let's take the example of Policy Analytics. The purpose is to determine new possible revenues through definition of new policy plans. In order to do so the service provider will need to analyze data usages, events and DPIs in light of the owned customer plan and services and definition of product and services portfolio.

At least the following data listed in the table below are needed:

Group of Data	Availability	Expandability	Resolution	Volume	Retention	Quality	Type of Infrastructure
DPI	via batch		maximum granularity	high	keep 6 months with high granularity for the entire customer base	Cured	Columnar DB
CDRs	via batch		maximum granularity	high	keep 6 months with high granularity for the entire customer base	Cured	Columnar DB
Events	via batch		maximum granularity	high	keep 6 months with high granularity for the entire customer	Cured	Columnar DB

Group of Data	Availability	Expandability	Resolution	Volume	Retention	Quality	Type of Infrastructure
					base		
Products Catalogue	via batch		only few attributes needed	low	all products which are in use for the 6 months period	Cured	Columnar DB
Plans and services subscribed by customer	via batch		maximum granularity	medium	all subscribed plan and services for the 6 months period for the entire customer base	Cured	Columnar DB

Let's take another example of automated top-up stimulation. The purpose is to determine the customers who are the most valuable in term of uplift revenue from long term perspective and deliver to them a marketing message suggesting to top-up at the most accurate moment. In order to do so the service provider will need to analyze the customer behavior over time and run some time series analysis against CDRs and usages, balance change history, top-up history and plans and services subscribed by customer.

From use cases perspective we need to analyze all these data through times series, compare them and correlate the trends to determine dynamically and overtime the segments of customers which could be good candidate to be targeted by contextual marketing campaigns. Such needs will lead to data scientist approach to discover pattern which are valuable from long term uplift revenue impact.

At least the following data listed in the table below are needed.

Group of Data	Availability	Expandability	Resolution	Volume	Retention	Quality	Type of Infrastructure
Balance change history	via batch		maximum granularity	high	keep 6 months with high granularity for the entire customer base	Uncured	Hadoop
Top-up history	via batch		maximum granularity	high	keep 6 months with high granularity for the entire customer	Uncured	Hadoop

Group of Data	Availability	Expandability	Resolution	Volume	Retention	Quality	Type of Infrastructure
					base		
CDRs	via batch		maximum granularity	high	keep 6 months with high granularity for the entire customer base	Uncured	Hadoop
Events	via batch		maximum granularity	high	keep 6 months with high granularity for the entire customer base	Uncured	Hadoop
Products from Product Catalogue	via batch		only few attributes needed	low	all products which are in use for the 6 months period	Uncured	Hadoop
Plans and services subscribed by customer	via batch		maximum granularity	medium	all subscribed plan and services for the 6 months period for the entire customer base	Uncured	Hadoop

From these 2 examples we can see that a service provider based on the same sort of data can have to handle them in 2 different paradigms to be used to address different use cases analytical processing purposes.

Example Step 6: Construct and Validate BDA Business Value Roadmap

Based on the analysis above, the following technologies were sought out and implemented.

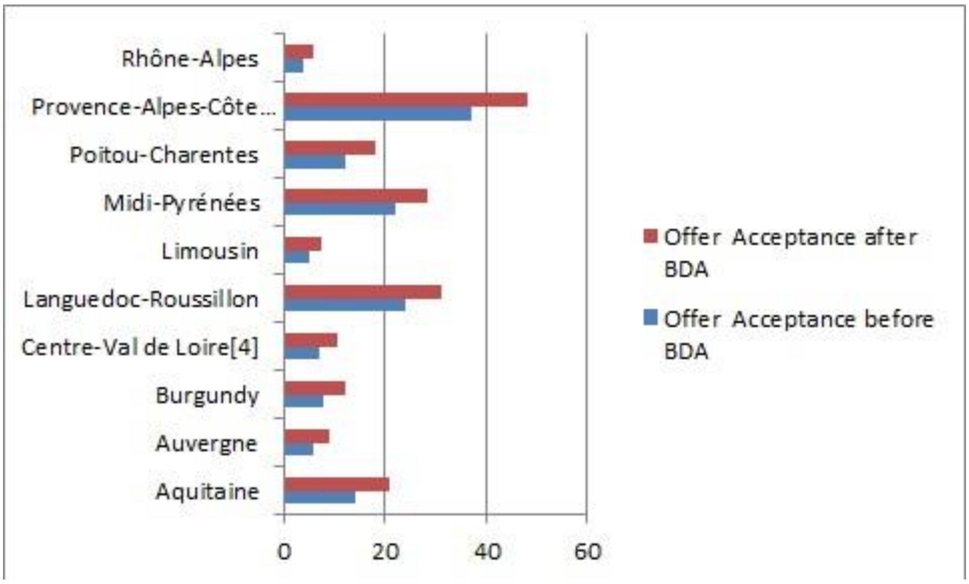
- Supervised Machine Learning algorithms
- Clustering algorithms

The information sources needed to implement the use case include

- real time and historical customer location information
- enrichment of customer social network information

With offers now better targeted to customers based on regional preferences, the overall offer acceptance was increased resulting in meeting the expected ROI. In addition NPS scores improved slightly.

Upon additional review of the results, it was decided that additional data sources and finer grained location information could improve results even further and the use case would be adapted to make these changes.



3. Big Data and Big Data Analytics

3.1. Big Data and Big Data Analytics

3.1.1. Big Data Definition

Quite a few attempts have been made on defining term Big Data and differentiate it from "regular" data. While every standard organization, consulting company, trade group may have a slightly different view on how to define the term, all of them refer to the characteristics of Big Data, commonly known as the Vs (three, four or more), not very far down into the definition section. At the time this Guidebook is published, the three Vs model, volume (amount of data), velocity (speed of data in and out), and variety (range of data types and sources), remains the dominating model to define the term Big Data.

A newer model (Big Data Paris, 2013) looks at Big Data as utilizing inductive statistics with data, the volume of which allows inferring laws and predicting to a certain extent future behaviors of the data. This is in contrast to traditional Business Intelligence, which uses descriptive statistics.

While three Vs may have described more comprehensively the characteristics of Big Data, we find the newer definition more relevant to Big Data Analytics, therefore more preferable to the work described in this Guidebook.

3.1.2. Big Data Analytics Techniques

No Matter which model one prefers, the value of the Big Data lies in the analysis results and the predictions/actions that derive from those results. The focus of this Project Group is not the Big Data itself, but the Big Data Analytics technologies and techniques that CSPs can use to unlock the values of Big Data in their business.

The analysis of Big Data requires extremely high performance against large data sets within reasonable response time. In order to satisfy these conditions, some "non-traditional" technologies have emerged during the past 10 years. Most, if not all the successful Big Data Analytics technologies excels in share nothing, massively parallel, scale out architectures, which are well suited for Big Data Analytics applications.

The following are some commonly known technologies used in Big Data Analytics applications.

MapReduce Framework and Hadoop

Published in 2004 by Google, this wildly successful framework is designed to efficiently process large volumes of data by connecting many commodity computers (cluster) together to work in parallel. MapReduce breaks the data flow into two phases, map phase and reduce phase. In map phase, chunks of data are processed in isolation by tasks called mappers. The outputs of these mappers are brought into tasks called reducers, which produce the final outputs. On the storage side, an underlying distributed file system splits large data files into chunks which are managed by different nodes in the cluster.

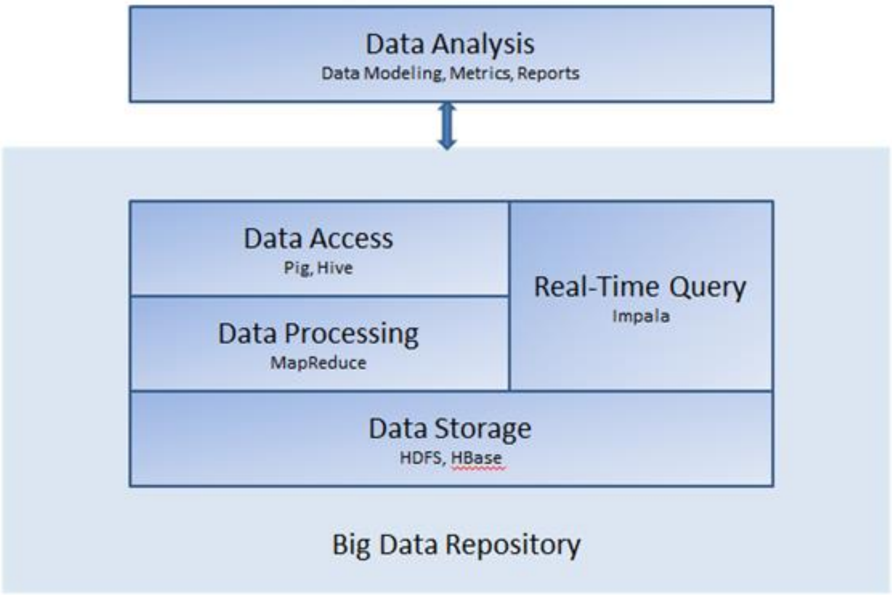


Figure 2: MapReduce Framework and Hadoop

Hadoop is the most well-known open source implementation of the MapReduce framework. The Hadoop Distributed File System (HDFS) provides storage support for the Hadoop Framework. HBase provides additional distributed database functionalities over HDFS. Data stored in HDFS are usually processed with MapReduce operations. Tools like Pig/Hive are developed on top of Hadoop framework to provide data access support over MapReduce/HDFS for upper-level analytics application. Newer tools like Impala bypasses MapReduce to provide real-time ad-hoc query capabilities. The Data Storage/Processing/Access functionalities provided by the Hadoop ecosystem are table stakes for a Big Data Repository.

NoSQL Store

NoSQL Stores provide highly available, scalable data storage systems with relaxed consistency guarantees compared to the traditional RDBMS. NoSQL Stores also provide flexible schema to allow heterogeneous columns on different rows of storage.

There are four different types of NoSQL database. First type is key/value store (e.g. Dynamo, Voldemold), which is inspired by the Amazon's Dynamo paper. The data model for this type of database is a collection of key/value pairs. This type of data store provides great flexibility for fast data storage in a programing environment.

Second type is column store (e.g. Cassendra, HBase). The data model is based on the original Google's BigTable paper. It stores data tables as columns of data rather than as rows of data. Because of this, it is well suited for Online Analytical Processing (OLAP), which typically involves smaller number of complex queries that access all the data stored.

Third type is document store (e.g. MongoDB). It is inspired by the database behind Lotus Notes. Document refers to a collection of information organized in a defined format (XML, JSON, BSON, MS Word etc.). Each document can be retrieved by a key (e.g. a path, a URL). Document store therefore is a collection of key-document pairs. Document store allows great flexibility storing semi-structured data and provides query facilities to retrieve documents based on their content.

Fourth type of NoSQL store is graph database (e.g. neo4j, Allegro graph). It is inspired by graph theory. Node, a key concept in graph database, is very similar to document in document store. Graph database stores key-node pairs, similar to key-document pairs. In addition, graph database adds relationships between the nodes and also stores key-relationship pairs. Graph database allows graph-based queries such as shortest path between two nodes and diameter of the graph.

Real-Time Query over HDFS

Although MapReduce framework is highly scalable for Big Data queries, it usually does not provide real-time responses needed interactive queries. Some solutions such as Impala attempt to solve the problem using a real-time ad-hoc SQL query processing engine directly over HDFS, bypassing the MapReduce processing to allow shorter response time. Additional optimizations such as compression can be used to accelerate response time further.

Solutions in this category provide a horizontally scalable implementation for real-time queries at Big Data level.

Search

In May/June 2013, MapR and Cloudera separately announced a new class of engines based directly/indirectly on Apache Solr/Lucene projects. The search feature of these engines allows text searches on data stored in HDFS, thus lower the technical expertise needed to perform Big Data Analytics. Real-life use cases for search include indexing assistance for semi-structured and unstructured data.

4. BDA Reference Model

4.1. Reference Model

The purpose of the reference model is to provide high level view of the functional components in a Big Data Analytics platform within a Big Data Solution ecosystem. By segregating layers of responsibilities between different functional components of the platform, we can get a clear view of the roles and responsibilities and lay the foundation for a common understanding of the Big Data Analytics domain.

4.1.1. Overview

This following diagram shows the high level overview of the Big Data ecosystem and specific functional layers of the platform. All layers provide external/internal APIs that serve both other layer functions and external third party applications in respective levels of data relevance and data density. The Reference Model allows highlighting the notions of 'data components'.

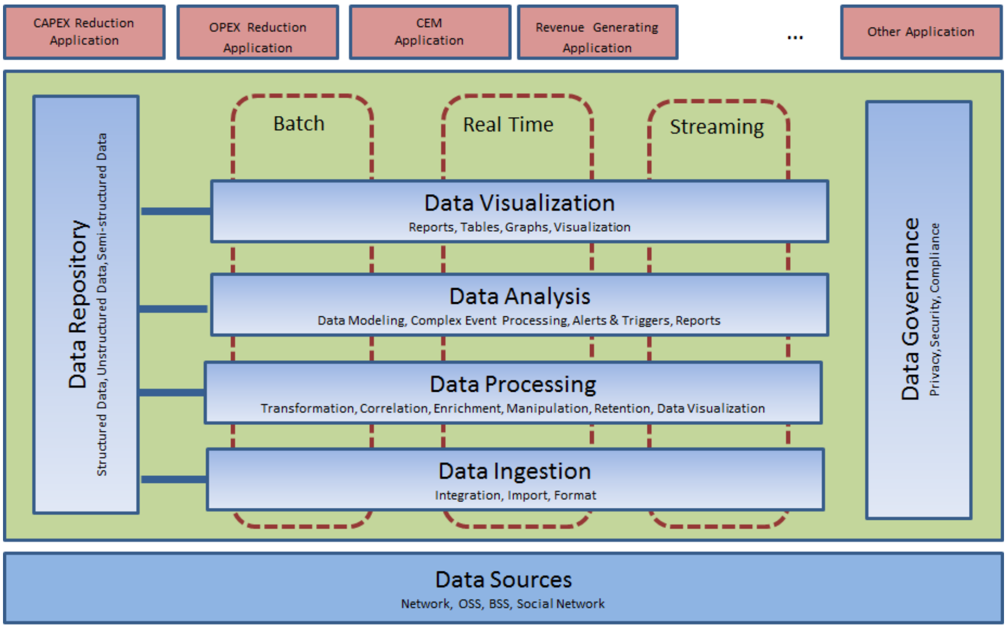

CAPEX Reduction Application
OPEX Reduction Application
CEM Application
Revenue Generating Application
...
Other Application
Batch
Real Time
Streaming
Data Repository
Structured Data, Unstructured Data, Semi-structured Data
Data Visualization
Reports, Tables, Graphs, Visualization
Data Analysis
Data Modeling, Complex Event Processing, Alerts & Triggers, Reports
Data Processing
Transformation, Correlation, Enrichment, Manipulation, Retention, Data Visualization
Data Ingestion
Integration, Import, Format
Data Governance
Privacy, Security, Compliance
Data Sources
Network, OSS, BSS, Social Network

Figure 3: Big Data Analytics Reference Model

Please note:

- The Reference Model diagram indicates the ‘total theoretical’ functionality that can ever be required by an arbitrary big data use case. Depending on the specifics of each use case, one may find that only a subset of this functionality may need to be involved. In this sense, all layers (with the possible exception of Data Ingestion) may be considered as optional when defining a big data use case.
- Layers have to be considered as an abstract grouping of similar functionality – not as architectural components of a particular big data platform. The actual mapping of the layers’ functionality on a particular big data platform may be left to the vendor’s discretion.

- Layers are not hierarchical neither sequential, like the ISO 7 layers model or the TCP/IP layers 4 model. With the exception of Data Ingestion which has always to be the layer accepting data from external data sources, all other layers can be sequenced in an arbitrary way, considered to be connected in a mesh-like way (all with all others). Please see the Data Flow sections for more details on the data flows between layers.
- The legend of the Big Data Analytics Reference Model is as follow
    - dark blue box: correspond to any data sources. They play a role as data provider but are not in a strict sense part of the big data analytics environment, other than data repository to take or receive data from, in a possibly pre-defined way. They will not be described in details until the description of concrete use cases where specific data sources may be required and then mentioned.
    - light blue boxes: correspond to an architectural component in which specific big data functions can take place.
    - red boxes: correspond to applications that leverage information in or produced by the big data analytics environment. They are the targeted "end-users" (in a broad sense).
    - red dashed boxes with round corners: correspond to the mode of ingestion of data and analysis that describes the way to load, manage, and analyze the data. 3 modes are considered here:
        1. 'Batch' mode: ‘Batch’ mode refers to off-line and planned processing. It starts on-demand and assumes that a huge memory space is available. It is a finite execution time program triggered by an external request that processes a finite set of data available at the request time.
        2. 'Streaming’ mode refers to on-line processing in which the analysis takes place over a pre-defined and continuously moving time window for all the analysis. It is a continuously running program that processes data flowing through. Streaming mode is related to Time-window-based Complex Event Processing. An example of it is for brand awareness, counting the number of tweets in twitter feed in which a specific brand is mentioned over a certain time window - typically between now and the last 10 to 30 mns (and how this evolves over time). Note that this time window could be delayed and does not strictly need an upper boundary of now (real-time). It may use historical or reference data (to compare computed KPIs with those used as references).
        3. 'Real-time' mode: refers to all other on-line processing, that are non moving time-window based. It is usually also based on (Complex) Event Processing. An example would be real-time location based marketing. There is no moving time windows required here. The only fact to be near a retail partner could trigger an adapted offer for/from this partner, with the communications service provider help.
        - Note: In ‘batch’ mode, signaling is split from data traffic whereas in ‘streaming’ mode, signaling is included in data traffic.

- The Big Data Repository can be thought of as an architectural component that, apart from storing raw or processed data, can (optionally) facilitate data flow between layers.
- Concerns regarding legal and regulatory compliance for consumer privacy often stifle the ability for CSPs to monetize data and form value-added data partnerships in the Data Value Chain. The privacy, security, and compliance functionality, within the Data Governance vertical, exists to address concerns via research-based data privacy preservation techniques Big Data Analytics application could be considered as a combination of the layers depicted on the Reference Model diagram.
- It was decided from Frameworx 16.0 to add so called ‘real time’ mode in order to put a special focus on real time decisions driven by data analytics in each CSP or DSP activities or business processes defined in eTOM: market/ sales, product, customer, service, resource, engaged party and enterprise. When either ‘batch’ or ‘streaming’ modes act on real-time, the related applications or services are supported by ‘real time’ mode in data analytics. ‘Streaming mode’ when executed in real time for example for learning or prediction use cases are then put under ‘real time mode’. In addition, it is admitted that any of these 3 mode does match to ‘data analytics’, ‘data visualization’, 'data processing', and 'data ingestion'. All these 'data components' can be implemented in any of the 3 modes: ‘batch’, ‘real time’ or ‘streaming’.
- The Reference Model can be also considered as supporting a BI, 'Business Intelligence', PaaS approach. More specifically, the Data processing and Data Analytics layers fully cover the BI functionality and can be used ad hoc by any external application or user interface, which can be both local and over the cloud (as a Service). In this frame, DaaS, standing for Data-as-a-Service can be considered.
- Generally speaking, as in the case of ‘data monetization’ use cases (please refer to document TM Forum IG 1338 in Frameworx 15.5), data can be accessible to third parties through APIs. Referring to ‘engaged parties’ such as in eTOM business processes, CSP/DSP and third parties involved in data monetization use cases are both referred as ‘engaged parties’.

4.1.2. Data Ingestion Layer

This layer is responsible for integrating with a variety of (big) data sources, importing the data into the big data platform and formatting the data into a uniform format. For Big Data, this layer is crucially important to handle the volume, velocity and variety of the data coming into the platform. This layer is where a number of optimizations can be implemented for Big Data. Functional modules in this layer should be inherently capable of scale out to accommodate the data input bandwidth and speed requirement

Key functionality supported by the Ingestion Layer is as follows.

Integration

Integrate with data sources and access data. This function is responsible for establishing connections with different systems, from which the data will flow.

Data Import

Import data from external data sources into the big data platform. Optionally, data can be labeled to denote the respective data source.

Data Formatting

Format data so that the same data from different sources has a uniform format before passing on to the next level application. This is essential for Big Data Analytics applications due to the variety of the data sources that may feed a single application. For example, IMSI from different 2G/3G/4G interfaces may have different encoding formats. This function in the Data Ingestion Layer will format IMSI to one single format throughout the rest of the layers.

4.1.3. Data Processing Layer

This layer accommodates a series of processing that can be applied on datasets ingested into the big data platform, such as transformation, correlation, enrichment manipulation as well as ensuring data quality and security. Such processing can be as follows.

Transformation

Map raw data into a data model in order to make data meaningful and usable. Typical data transformations include function categories, such as:

- Comparison
- Date & Time
- Logical
- Math
- Statistical
- Text
- Trigonometry
- Encoding
- List Management
- URL management

Correlation

Associate different representations/and data collected from various sources of the same business entity.

For example, this layer can associate the MSISDN taken from CDRs with the Customer ID taken from the CRM. Both numbers represent the same business entity, customer. Data collected from both sources can be correlated together to provide a richer set of information related to the customer.

Enrichment

Combine multiple data sources that refer to the same business entity (e.g. customer) in order to create a more complete view of the entity. In some cases, enrichment data sources can be from CSPs' various customer information databases. In some other cases, some enrichment data can be from the Big Data Analytics results.

For example, based on a customer's browsing history and locations, it may be inferred with high degree of confidence of the customer's gender, age, educational level, income level etc.

Dataset Manipulation

Functionality that can be applied to entire datasets, like:

- Union
- Intersection

- Sorting
- Filtering
- Compression
- De-duplication/Duplication
- Group Series functions
- Aggregation functions

Data Quality Assurance

Perform the following functions to assure data quality:

- Data cleansing
- Data integrity assurance

For example, data with checksum errors may be logged and thrown away.

4.1.4. Data Analysis Layer

This layer supports advanced Big Data Analytics either in batch, streaming and in real-time modes by supporting functionalities such as calculation of metrics, data modeling, CEP and machine learning.

Data Analysis layer relies on a number of techniques, including:

- Event-pattern detection
- Learning in real-time
- Event abstraction
- Modeling event hierarchies
- Detecting relationships (such as causality, membership or timing) between events
- Abstracting event-driven processes
- Generation of alerts/triggers to action.

The key functionalities of Data Analysis layer are as follows:

Descriptive/Predictive/Prescriptive Modeling

Conduct Descriptive/Predictive/Prescriptive Modeling (explaining the past / predicting the future/recommending next best action) by utilizing Machine Learning / Data Mining algorithms, such as:

- Classification
- Clustering
- Pattern Mining
- Recommenders / Collaborative Filtering
- Statistical Relational Learning
- Text/Speech/Video Analytics

Complex Event Processing

Most Complex Event Processor (CEP) solutions and concepts can be classified into two main categories:

- A Computation-oriented CEP solution is focused on executing on-line algorithms as a response to event data entering the system. A simple example is to continuously calculate an average based in data on the inbound events.
- A Detection-oriented CEP solution is focused on detecting combinations of events called events patterns or situations. A simple example of detecting a situation is to look for a specific sequence of events.

CEP offers the functionality that makes it possible to implement Big Data Analytics scenarios which require real-time processing. More specifically, CEP controls the processing of streaming data, the correlation of occurring events and the calculation of KPIs on an ongoing basis. Driven by user-supplied business rules, CEP generates alerts or triggers for subsequent actions by external systems.

In the context of big data, CEP can be implemented by massively parallel-enabled complex event processors like, for example, Twitter's open source Storm (http://storm-project.net/).

Generation of Alerts/Triggers to Actions

The outcomes produced by Data analysis can trigger alerts and actions.

- **Alerts** are mainly destined to humans for further consideration (M2H)
- **Triggers** are mainly destined to other applications or systems that automatically proceed to the corresponding actions (M2M).

For example, a network performance monitoring application may use CEP to monitor the alarms from network elements. When the alarm number/severity exceeds certain thresholds, the application will generate a critical alarm to the network operator and trigger the policy changes to re-route network traffic away from the affected subset of the network.

Metrics Calculation

Calculate relevant business metrics, such as:

- TM Forum's Business Metrics (e.g. Frameworx Metrics, Customer Experience Management Index, Balanced Scorecard, etc.)
- Arbitrary, ad hoc metrics.

Reports Generation

Data Reports can be generated in real-time, on a daily/weekly/monthly basis, or on-demand. They can be used to visualize the Big Data Analytics results. There are a lot of visualization tools off the shelf that display efficiently these data reports.

4.1.5. Data Visualization

The primary goal of Data Visualization is to communicate information clearly and efficiently to users via statistical graphs, plots, information graphics, tables, charts, heat maps and info graphic. This communication has to be done with enhanced user experience. Effective visualization helps users analyze and reason about data and evidence. It makes complex data more accessible, understandable and usable. Tables are generally used where users will look-up a specific measure of a variable, while charts of various types are used to show patterns or relationships in the data for one or more variables.

By extension, 'data visualization' may as well refer to 3D representation or to 'augmented reality' graphic representation.

4.2. Big Data Repository Vertical

This layer provides storage of all data within the big data platform which can be either in the original 'raw' form in which it was ingested into the system or in any intermediate, processed form produced by any other of the Reference Model layers.

The Data Repository is typically a data store that provides scaling and flexibility needed to handle the data volume and can be:

- Local, e.g. within a CSP's internal data center, or
- Accessed over a private or public Cloud (or multi-cloud).

The Big Data Repository interacts with all other layers and can be thought of as the equivalent to a 'data bus'.

Big Data Repository can store various types of data: either unstructured, structured or semi-structured.

The concept of unique analytics big data repository, standing for unique ABDR, as defined in GB979 addendum D, provides a way to handle data, by multiple and heterogeneous data analytics platforms, various business entities, various 'engaged parties' (as defined in eTOM), various affiliated companies within a same corporate Group. By having such unique ABDR, that breaks the silos, unique ABDR, associated with Data Governance, any CSP/DSP can achieve digital transformation by managing data as enabler of digital transformation.

Data is standardized with metadata: these standardized data are called 'data entities'. Then all actors in Data analytics area can use the same language: CSP, DSP; technology providers, 'engaged parties'.

4.2.1. Unstructured Data

It refers to data that either does not have a pre-defined data model or is not organized in a pre-defined manner. Unstructured information is typically text-heavy, but may contain data such as dates, numbers, and facts as well. This results in irregularities and ambiguities that make it difficult to understand using traditional computer programs as compared to data stored in fielded form in databases or annotated (semantically tagged) in documents.

4.2.2. Structured Data

It refers to data that is organized in a structure according to a pre-defined data model. Structured data is also searchable by data type within content. Structured data is understood by computers and is also efficiently organized for human readers.

4.2.3. Semi-Structured Data

It is a form of structured data that does not conform to the formal structure of data models associated with relational databases or other forms of data tables, but nonetheless contains tags or other markers to separate semantic elements and enforce hierarchies of records and fields within the data.

It is important to understand that in different Big Data Analytics platform deployment scenarios, it is rarely the case that Big Data Repository is a single technology entity. Rather, it is more likely to be a collection of different storage/processing technologies stringed together during implementation/optimization, so that specific technologies can be used to cache, aggregate, optimize the data so that they can be easily and effectively processed by other functional modules in different layers in the model for specific Big Data Analytics use cases.

It should be stressed that, the breakthroughs created by the evolution of big data storage technologies (like HDFS) make storage a commodity. As a result, contrary to the traditional data warehouses where data was a scarce/precious asset, big data platforms now have the 'luxury' of storing both raw and intermediate data that, in turn, can be both historical and streaming. This justifies the term 'big' in big data and multiplies the value that can be extracted by processing the entire dataset vs. only samples of it.

One typical usage of Big Data Repository today is to use NoSQL DB with massive parallel processing such as MapReduce to quickly store the big data. Then a subset of the data can be queried or exported into a traditional RDBMS for BI/Reporting tools to process and present results.

Any necessary visualization takes data stored in the Repository and, with the use of graphical tools, produces reports, dashboards, scorecards, infographics – potentially interactive.

4.3. Data Governance Vertical

4.3.1. Overview

Big Data is not only a challenge for Enterprise Organizations and Governments when dealing with their customers or citizens but there is also a challenge for the security domain as well; new tools and methodologies will need to be embraced to deal with this emerging threat. In the recent past we have seen a vast increase in security measures and this has led to an exponential increase in the collection and analysis of increasingly larger amounts of event and security contextual data. Building customer trust is of utmost concern for CSPs to improve their bottom line through monetization of their Big Data Analytics solution as mentioned in a recent TM Forum research report.[1] In order to build customer trust, CSPs should:

- Adopt and consistently adhere to industry best practices and codes of conduct; and
- Comply with jurisdictional laws and regulations.

These increases are forcing the industry to consider an integrated approach to Big Data which encompasses Privacy and Security as constituent parts of the best practice model and forms part of the Big Data Analytics Guidebook.

In the same global scope, privacy concerns are as well a challenge to be dealt by CSPs willing to benefit from Big Data approach.

A whole Data ecosystem has to be put in place in order to exchange data amongst CSPs, public sources (open data), other players that take advantage of data provided by CSPs, and data originating from other industry verticals that benefit from crossing with CSPs data. In particular, the data owner, who is accountable for customer data is one of the key actors in the data value chain.

Moreover inside CSP organization, data governance is a global business process. Within some CSPs, data governance is implemented through a data governance Board. Data governance covers the areas of security, privacy and compliance to legal and regulatory jurisdictions. Data governance defines the policies to be applied to each category of customer or network data and the induced rules to be enforced.

In addition, Data governance is an umbrella term, which encompasses various functionalities of the Reference Model.

Compliance with jurisdictional laws and regulations and with contractual obligations has to be handled by data governance Board and the results in terms of rules enforcement is visible in the 'data governance' considered as one of the application component of the Big Data Analytics Reference Model.

Data governance processes

Data governance, DG, processes are defined in document TR261and will be integrated in eTOM Fx 16.5, as level 2 processes in 'enterprise' domain.

The 6 main processes are as follows:

- Define DG strategy
- Define DG organization
- Define DG policy & processes
- Define DG measurement & monitoring method
- Select DG technology
- Design continuous improvement strategy

These 6 processes are coordinated by using DG roadmap.

DG organization can be split in DG council and in each business unit, data stewards. DG council needs to be driven by executive leadership, to become data-driven organization deals with changing the way people work and it has to be done from top down approach but also to win adoption by all positions people within the company.

Data ownership is assigned to the business entities.

Reference Model Functions

The Data Governance Layer encapsulates all other layers of the Big Data Analytics platform to address the best practices introduced above and provides the following functions:

- **Privacy**: Management, Protection & Preservation
- **Security**: Encryption, Authentication, and Access Control
- **Compliance**: Legal and Regulatory

Contextual Privacy Dimensions

These functions address the privacy of personal data along the following contextual dimensions:

- **Collection** – the acquiring of a Customer's personal data
- **Use** – the storage, manipulation, and application a Customer's personal data applicable to the CSP's business
- **Disclosure** – the release of Customer personal data or any aggregate that can be linked back to an individual Customer or End User.

4.3.2. Privacy Management

Privacy Management addresses the Customer's need for transparency, choice, and control by providing visibility and configuration of privacy preferences and practices.

Privacy applies to Personal Identification Information (PII) data. Anonymization techniques can transform PII data into non-PII data.

CSP-Managed Privacy Policy

- Default privacy policy
- Collection, use, and disclosure context
- opt-in/opt-out preferences
- visibility into the collection, use, and disclosure personal data
- data retention limits

Customer-Managed Preferences

Privacy Protection

Privacy protection provides privacy assurance for an individual Customer's data via adherence to:

- CSP-managed privacy policy and
- individual Customer-managed preferences

Protection is provided in the security function to control the collection, storage, use, and disclosure of Customer data based on the above policy.

An alternative approach is for CSPs to provide their customers with a 'privacy management dashboard' allowing customers/end-users to control precisely how their personal data will be used. It would be better still if such a dashboard were not tied to any specific CSP or supplier; i.e. would work with all.

Techniques such as 'data analytics privacy risk assessment' as defined jointly by the TM Forum data analytics and Privacy Groups would help shape the privacy dashboard.

Privacy Preservation

Privacy preservation addresses concerns regarding the disclosure of Customer data, at the individual and at the aggregate level of granularity.

Privacy preservation methods should be research-based and should provide an acceptable balance between privacy preservation and data utility.

Anonymization Techniques

Privacy preservation may be linked to anonymization techniques that are required to prior to disclosure of customer data, such as:

- k-anonymity;
- pseudonymization; and
- redaction of Personally Identifiable Information (PII).

Differential Privacy Techniques

Recent research, in the areas of differential privacy techniques, offers mathematically proven privacy-preservation techniques.[2] Differential privacy seeks to preserve privacy (minimize risk) while maintaining data utility. Differential privacy does not guarantee perfect privacy or perfect utility.[3] While, differential privacy is not perfect it is general more acceptable than PII redaction or pseudonymization when re-identification of a single individual is of concern.

Synthetic data set to provide CSP anonymized data to third parties

Open data – safe data sharing - to provide Call Data Records (CDR) in synthetic data set format to enable the provision of anonymous data.

There is a technique called 'synthetic data set format' to provide CDRs (Call Detail Records) as anonymized data to third parties. This technique uses both anonymization and differential privacy techniques.

In order to disconnect data and real mobile phone user, Orange have proposed the creation of a synthetic dataset from the original CDRs, namely a fictitious set of data having the same characteristics as a set of true data, in such a way that it is not possible to re-identify the initial true users. Such method is still at the experimental stage but initial results looks really a promising one.

Details about this technique are found in section 9 of this document.

4.3.3. Security

Encryption

Encryption capabilities provided for data at-rest and data in-transit. Stored data may reside within the Big Data Repository or in temporary storage locations within Data Ingestion and/or Data Exchange Services. Data in-transit should be encrypted for all access (e.g. Data Ingestion, Date Exchange Services, and applications).

An efficient encryption system requires at the very least a double set of keys otherwise it is not resilient.

Authentication

Context-aware authentication services are provided for all layers that allow external access to the Big Data Solution. A context includes the accessed Layer, the role of the authenticated entity, and the purpose of the accessing application.

Access Control

Access control administration capabilities are provided for both the CSP and the Customer. The CSP defines role-based and context-aware access control per their privacy policy and security best practices.

Access control provides a reasonable means for a Customer to access their personal data and Privacy Management configuration.

4.3.4. Compliance

Compliance is the adherence to jurisdictional laws and regulations according to the context meaning mainly location, date and time in which the data is collected, used, and disclosed. Compliance may be required (e.g. laws or regulations) or optional (e.g. best practices, codes of conduct).

Compliance deals with the possible certifying entity in charge of dealing with data governance inside a country and then inside a CSP. As set up in Germany, it may be a Trustee, an independent entity.

Legal Compliance

Legal compliance refers to adhering to jurisdictional laws - country-based, state-based, or federation-based (e.g. multi-country) - that apply to customer data in the country or in the state it belongs to.

An example of multi-countries laws is the European legal framework with the EU, European Union, directive related to privacy called ePrivacy directive.

And what will come into force at EU, in 2018 is the global GPDR, Global Data Privacy Regulation framework.

For instance, in some countries, customer data must be kept inside the country or has to be traceable.

Regulatory Compliance

Regulatory compliance is handled by country-based or state-based entities and by multi-countries entities or federation-based entities.

For instance, the French National Commission on Computing and Liberty (CNIL) is responsible for protecting the human rights and the identity and privacy of individuals. Prior to any data collection, an authorization has to be allocated by this entity to the CSP or DSP.

The duration of data collection is limited to 1 year or 14 months as for Call Data Records.

Best Practices / Codes of Conduct

In addition to jurisdictional laws and regulations, CSPs may also implement codes of conduct and/or best practices as recommended by industry groups, non-legislative government entities. For example, the White House proposed the *Consumer Privacy Bill of Rights* which addresses the protection of personal data, “including aggregations of data, linkable to a specific individual.[4] ” The Mobile Marketing Association (MMA) provides a *Global Code of Conduct* which addresses privacy standards for mobile device users.[5]

5. Analytics Big Data Repository

5.1. Analytics Big Data Repository

The ABDR is CSP's data repository used mainly for analytical purposes, permitting efficient and straight forward re-use of data for multiple purposes. The ABDR specs are defined by the TM Forum. This permits creating standardized implementations across operators, and straightforward reusability of the data by ABDR compliant solutions.

Informally, an ABDR implementation is a collection of unique multiple independent data entities that have a clear definition, aka a data dictionary, per data entity. For example an ABDR can include CDRs, DPI, and billing records and a data dictionary per each of these record types. The data dictionaries and data entities are not arbitrary but are according to the ABDR definition by the TM Forum.

Let's look at the 3 descriptive elements, ***unique multiple independent*** data entities.

Unique – denotes that the same record should reside only once in the ABDR, and should not be replicated. E.g., if we collect CDRs from the switch the CDR collected from the switch will reside only once in the ABDR. Note that if we collect for the same event both a CDR from the switch, a post-mediation CDR, and a post-rating rated CDR, these are 3 separated records, and not three copies of the same record, therefore all the three records can reside in an ABDR implementation. Replication is allowed in the ABDR for purposes of high-availability, performance, and disaster recovery. This allowed replications should be transparent to the applications using the ADBR

Multiple – denotes that many different data entities reside in the ABDR.

Independent – independent or loosely coupled, the ABDR does not impose relations between the different data entities. E.g., there is an entity of a subscriber and an entity of a CDR, obviously there is an implicit relation between these entities, e.g., the calling number might belong to a subscriber. However the ABDR definition does not capture or imposes this relation. In the ABDR each entity is isolated, applications can implement different relations between entities on top of the ABDR, include full information models like the SID.

Typically the ABDR will serve as a data repository of the analytical systems, and not of the operational systems.

The data in the ABDR will typically include "near-raw" data in a simplified format, e.g. simplified switch CDRs, and it will also include data generated as result of analytical processes, e.g., churn scores. An example of "near-raw" data are switch CDRs, switch CDRs come in many shapes and flavors, putting them as is in the ABDR will require each application to do a complex mediation process, unnecessarily increasing the burden and cost per application, hence in the ABDR we define a simplified "near-raw" switch CDR format, it will be a text and not a binary format, it will include a set of predefined fields, such as calling number and called number, yet at the same time it will the data will be kept "near raw", in the sense that the information will not be enriched "too much" with external information, and will not cleaned "too-much", e.g. CDRs with errors or with missing information will not be dropped.

The ABDR does not dictate a certain implementation technology; however, it is clear that today's large ABDR data repository implementations will have a Hadoop and HDFS component and in many cases other technologies will be combined, e.g., columnar data bases, non-sql data bases, file systems, and even Enterprise Data Warehouse

Even though the ABDR concept can and should be used in many industries, the ABDR is defined by the data entities and data dictionaries it contains. For CSPs the entities and their dictionary are defined by the TM Forum. In the future there will be multiple releases of the ABDR definition.

6. Data Flow

6.1. Data Flow

6.1.1. Batch Processing Data Flows

Batch Processing is a common solution model for data warehouse/data mining type of analytics. Data from different sources are collected over time and imported into a big data analytics platform. Then a data mining application will look through the data stored in the platform, apply specific analytics logic, and produce reports and/or any other mode of visualization. This model of processing is well suited for applications that need a large amount of data to produce accurate, meaningful intelligence. Typically time is not a pressing factor in consideration. Moreover, the specific type of analytical operations may not be fully understood nor anticipated at collection time. Therefore the main concern is to have as much data stored as possible for post-processing at a later time.

Offline Batch Processing models provide tradeoffs between data availability and data storage size limits. The more data available for later processing, the more storage it needs to keep these data. It also provides tradeoffs between analytics accuracy and reporting time. The more thorough the analytics computation, the longer it takes to get the results.

The solid line data flows represented in the diagram below correspond to the Offline Batch Processing Solution Model, as described in Section 6.2 of this document. This is typical of data warehouse or mining type of analytics systems, where data is ingested in batch files. The data from these files are then formatted and enriched, based on which analytical results can be derived. Transfer of data between the different layers is represented below.

Data Source to Data Ingestion

This data flow represents the communication between the Data Ingestion Layer and a variety of big data sources. The data is imported in batch files, which are typically in binary or ASCII. Data transfer between the two layers can happen via the following modes:

Pull: Where the big data analytics platform polls different data sources and extracts files to process. E.g. the analytics platform polls the CRM system periodically to read details of customers who have recently undergone a plan change.

Push: Where the data sources FTP data files to the big data analytics platform for processing. E.g. a billing system pushes usage data in periodic intervals to the big data analytics platform.

Some of the key mechanisms that are used to ingest data are mentioned below:

- **ETL** (Extract-Transform-Load) involves
    - Extracting data from outside sources
    - Transforming it to fit operational needs, which can include quality levels
    - Loading it into the end target (big data repository)

- **ELT** (Extract-Load-Transform) is a mechanism where the processing is performed at the database level, thereby enhancing performance in big data analytics deployments.
- **ETLT** (Extract-Transform-Load-Transform) comprises of a mixed workload between ETL and ELT, since different mechanisms might be needed in different scenarios. ETLT optimizes performance, solves complex business logic and makes it simpler to cleanse data.

Data Ingestion to Data Processing

The files imported by the Data Ingestion Layer, as mentioned above, are unformatted when they are received. This data is formatted by the Ingestion Layer, after which they may be stored in the Big Data Repository.

Alternatively, as this data flow represents, the formatted data files can be read by the Data Processing Layer in order to perform the next level of processing, namely, transformation, correlation and enrichment. E.g. considering the scenario above, the file (containing customer data who have undergone a plan change) is matched with other data (like, plan launch date, as available in the product catalogue), thereby resulting in an enriched set of information containing customer and plan details.

Data Processing to Data Analysis

The meaningful and enriched data, as output by the Data Processing Layer can be stored in the Big Data Repository.

Also, metrics and reports can be created directly by reading the processed files from the Data Processing Layer. This is represented using the data flow between the Data processing Layer and the Data Visualization Layer. E.g. considering the scenario above, the file (now containing information like customer ID, plan A, plan B, plan A launch date, plan B launch date, etc.) can be analyzed and metrics/reports generated that would help determine the average duration for which plan A was used by the operator’s customer base before they switched to some other plan, which hence would give an indication of plan A’s success rate.

Data Ingestion to Data Analysis

In certain cases, the data imported by the Data Ingestion Layer might already be sufficiently enriched in advance, and might not require further processing in order to derive metrics and reports out of them. E.g. batch file feeds containing billed and unbilled usage from a billing system to the big data analytics platform. The data present in such files are formatted, and contain sufficient information for the Data Analysis Layer to generate metrics and reports.

This is represented in the data flow between the Data Ingestion Layer to the Data Analysis Layer, where no further transformation of data is required, and the ingested data can be used directly for performing further analysis.

6.1.2. Real time and Stream Processing Data Flows

In Real time and in Stream Processing solution model, the data is stored as it comes into the analytics system. It is then processed either instantaneously or in time windows. As a result, the analytics system can provide reports in near-real-time on the data already received and actions can be taken immediately based on the reports. Also, actions can be taken in real-time based on events as and when they occur in a source system.

Although some systems use this model to improve overall performance/response time for their data mining analytics, a more common use of this model is to satisfy some real-time use cases such as location-based marketing and fraud detection/prevention.

While this is a powerful solution model that addresses some of the concerns for Offline Batch Processing, the analytics result can be skewed by the limited dataset it processes in each window.

The data flows represented here correspond to the Real-time or Stream Processing Solution Model, as described in Section 3.5 offline of this document. This characterizes data which is received in online/real-time mode by the big data analytics platform for immediate processing. The focus of this model is to provide accurate analytical results based on current data available. The data may or may not be stored for future use. Transfer of data between the different layers is represented below.

The instantiation of such data is usually in the form of messages or API calls to the big data analytics system, and is referred to as an “event”.

Data Source to Data Ingestion

This data flow represents the transfer of events between the Data Ingestion layer and a variety of big data sources. This can happen via the following modes:

Pull: Where the data analytics platform polls data sources for available unprocessed events, and retrieves the same for processing. E.g. In the case of certain location-based offers, the analytics engine reads the subscriber location information at frequent intervals, and generates offers based on the same.

Push: Where events are sent for analysis as and when they occur in the source systems. Hence, the direction of data transfer is from the source system to the data analytics platform. E.g. While a customer is browsing products/services in a web portal of the service provider, events containing web page details are sent to the analytics platform, based on which real-time recommendations can be generated.

Some of the key mechanisms that are used to ingest data are mentioned below:

- **Data Streaming**, in which a sequence of data packets are transmitted from the data source to the data analytics platform
- **Replication**, in which changes in the source data are sent to the analytics platform, which is then processed.
- **Microbatch ETL**, which is like traditional ETL, but happens in frequent intervals.

Data Ingestion to Data Processing

The event which is read by the Data Ingestion Layer, as mentioned above, contains raw data. After being further processed, this may be stored in the Big Data Repository.

Alternatively, as this data flow represents, the event is then transferred to the Data Processing Layer in order to perform the following level of processing, namely, transformation, correlation and enrichment. E.g. following the scenario mentioned above, the event containing products/services currently being viewed by the customer is correlated with customer demographics information (age, gender, etc.), thereby enriching the event further..

Data Processing to Data Analysis

The meaningful and enriched event, as output by the Data Management Layer can be stored in the Big Data Repository.

Also, this information can be used to generate analytical results in real-time, by transferring the event to the Data Analysis Layer, as represented in this data flow. E.g. considering the scenario mentioned above, the enriched event is then used to come up with a set of recommendations that can be made to the customer.

Data Analysis to Complex Event Processing

The data generated by the Data Analysis Layer is transferred to the Complex Event Processing Layer, in order to perform complex computations and send alerts to the end customer. E.g. based on recommendations, as generated by the Data Analysis Layer, CEP calculations are done and based on other factors (like propensity analysis), the most relevant offer is sent to the customer in real-time.

Data Ingestion to Data Analysis

The raw event that is read by the Data Ingestion Layer from data sources might not need further enrichment, and hence can be transferred directly to the Data Analysis Layer, as represented by this data flow.

E.g. during checkout from a shopping portal, an event would contain all necessary information for the analytics engine to generate personalized and targeted up-sell and cross-sell offers based on the products that the customer currently has in their basket. Hence this event does not need any further enrichment, and can be directly transferred to the Data Analysis Layer.

Data Processing to Complex Event Processing

This data flow represents transferring an event directly from the Data Processing Layer to the Complex Event Processing Layer, which can then be subject to further complex processing.

E.g. during a live interaction with a CSR, 'Customer Service Representative', the information of the customer, as output by the Data processing Layer can be directly used for further complex event processing, like detecting the call pattern and generating offers or alerts based on the same.

6.1.3. Data visualization

The functional component called 'data visualization' is independent from the type of data flows: whether batch or real time/stream modes. Hence there is only one 'data visualization' function in the illustrating graph below.

6.1.4. Flow graph: illustration as an example

Data traffic processed either by batch or real time/stream modes may come either from the same data sources or from different data sources.

For example, data may be ingested in stream mode and then both stored in data repository for batch processing to come through data processing and data analysis layers and as well follows the data processing and data analysis layers in real time/stream mode.

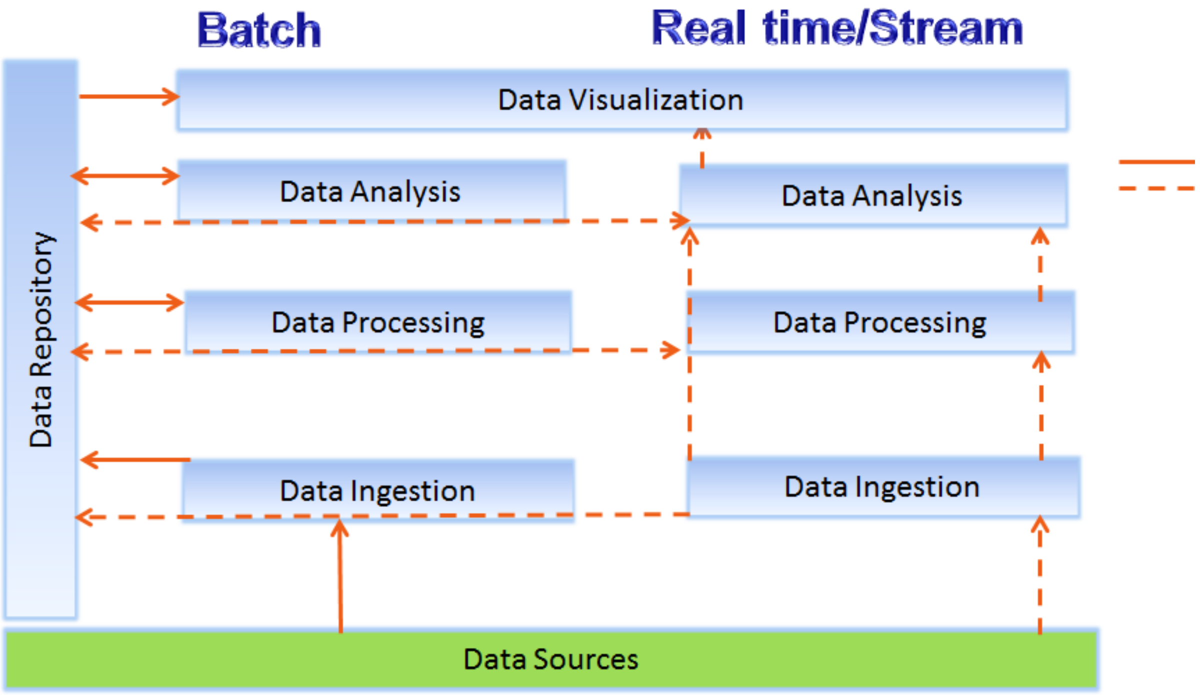
The results of data analysis layer in real time/stream mode may be stored in data repository and then used in batch mode processing.

There may be several exchanges of data traffic between batch and real time/stream modes.

Data Traffic

The 'orange' colored lines represent data traffic in the diagram below.

The solid lines represent batch mode processing while the broken lines depict real-time/streaming mode.


Batch
Real time/Stream
Data Visualization
Data Analysis
Data Analysis
Data Processing
Data Processing
Data Ingestion
Data Ingestion
Data Repository
Data Sources

7. Additional Details on Synthetic Data Model

7.1. Additional Details on Synthetic Data Model

Models of human mobility and communication have wide applicability to infrastructure and resource planning, analysis of infectious disease dynamics, ecology, administration, politics and more. The abundance of spatiotemporal data from cellular telephone networks affords new opportunities to construct such models. Furthermore, such data can be gathered with greater detail at larger scale and lower cost than traditional methods, such as transportation or census surveys.

The most popular available datasets come from operators' billing systems, the Call Detail Records (CDR) which have largely been used in the research or for the company's marketing or business intelligence purposes. Even the simplest phone leaves behind extensive CDRs that are preserved by mobile carriers. These records — on the time a voice call or text message was placed, and the identity and location of the cell tower involved — give the approximate locations of the phone's owner. Over time, they can be used to develop an accurate trace of the user's mobility.

Operators are extremely cautious about sharing their customers' data for the evident legal, business and privacy reasons, but some research work has been done in the one-to-one strict contracting framework between operators R&D (ex. AT&T, Orange, Telefonica, Telenor…) and the selected research teams. The proven value of those digital traces for the behavioral research has created a strong demand from researchers in complex networks, transportation, epidemiology or urban planning as well as other public and private institutions. To partially respond to this demand, the Data for Development, D4D, scientific challenge was organized in 2012 by Orange with CDR data from Orange Ivory Coast (other initiatives have also emerged, e.g., Telekom Italia Big Data Challenge 2014, D4D Senegal 2015 by Orange with data from Orange Senegal, please refer to the Orange Data4Development site for more information); large CDR samples were provided to bigger, but still preselected, number of research teams. The popularity of these initiatives, and new perspectives for research, NGO or health actions, illustrate the growing public value from telecom data sharing, in a secure and privacy compliant manner.

There are several proposals for the provision of data in a safe manner that is compliant with national privacy legislation when the operator penetration rates are important on the national level.

Providing a third party with communication data, such as exchanges between customers, or a communication graph (who calls who) is clearly too risky and privacy breaches would be inevitable. .An alternative is to exploit aggregated data, e.g.: antenna to antenna exchanges by hour, by day, etc. This data can provide an insight into communication dynamics across a large territory (country, region…) and provide an opportunity to identify and analyze the social structure of communities within it.

However it is not straightforward. When using cell location as a proxy for movement in human mobility analysis, a number of procedures have been developed to provide researchers with data samples. Temporal sampling is supposed to protect privacy as every X days a new random sample is drawn and the possibility of customer identification is lower. However some recent work showed that even poor mobile location and a short time period can open the door to data re-identification (cf. Yves-Alexandre de Montjoye, César A Hidalgo, Michel Verleysen, Vincent D Blondel, Unique in the Crowd: The privacy bounds of human mobility, Scientific Reports 3, 1376, 2013).

In order to disconnect data and real mobile phone user, Orange have proposed the creation a synthetic dataset from the original CDRs, namely a fictitious set of data having the same characteristics as a set of true data, in such a way that it is not possible to re-identify the initial true users. The method is still at the experimental stage but initial results look promising.

The main constraint regarding this technique is that one has to determine, at the start of the process, the set of useful properties that should be maintained. It is therefore essential to know how the synthetic data will be used in the future. The benefits of such technique is that the resulting data set is a good compromise between privacy and usability.

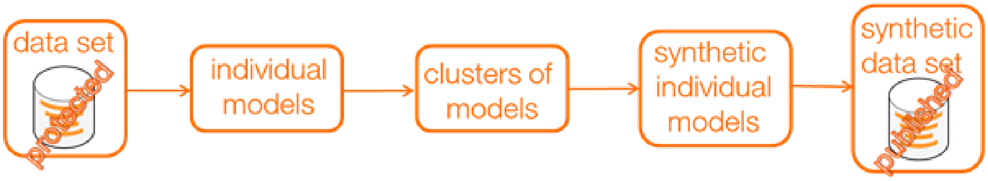
The process of creating synthetic data is complex but can generally be broken into a series of several steps:

- **Construction of the individual models from CDRs.** First, for each individual in the original dataset, an individual model is built, which represents in a compact manner the mobility and the call behavior of this user. This step, in the case of CDRs, comprises computing for each representative user inside the initial database, the number of visited cell towers, the entropy of the visited cell towers, the number of calls, the time between two calls, the covered distance per day, etc.

- **Clustering of the individual models.** These individual models are then grouped by clusters, namely groups of similar models. The objective of the clustering algorithm is to group the individual models that are close in the same cluster (intra-similarity) while putting the models that are different in separated clusters (inter-dissimilarity). To detect such (dis)similarities, this clustering is guided by a distance metric between models (small distance corresponds to quasi identical individual models, while a large value indicates a significant difference between the behaviors captured by these models). To protect the privacy of initial users, each cluster is crafted such that it contains at least k different models of users (in order to ensure a form of k-anonymity, where k is a predefined parameter given as input to the sanitization method). In addition, a minimal level of diversity is maintained in each cluster to avoid the situation in which all the models contained in a cluster are almost identical to each other.

- **Generation of synthetic individual models.** Afterwards, the individual models belonging to a cluster can be aggregated in a representative model summarizing the mobility and call behavior of the group. These representative models can then be used to instantiate the individual mobility models for artificial users. In a nutshell, the algorithm creates synthetic individual models.

- **Generation of synthetic CDRs.** Finally, using these models themselves, the algorithm samples the representative models in order to produce a dataset of synthetic CDRs whose global characteristics are close to the ones of the original CDRs.



The resulting synthetic CDRs are then closely related to the individual models that have been created and can hardly be used for other purposes. They can be published as created, without compromising the privacy of individuals whose personal data was embedded in the original CDRs.

8. Administrative Appendix GB979 R16.5.0

This Appendix provides additional background material about the TM Forum and this document. In general, sections may be included or omitted as desired; however, a Document History must always be included.

8.1. About this document

This is a TM Forum Guidebook. The guidebook format is used when:

- The document lays out a ‘core’ part of TM Forum’s approach to automating business processes. Such guidebooks would include the Telecom Operations Map and the Technology Integration Map, but not the detailed specifications that are developed in support of the approach.
- Information about TM Forum policy, or goals or programs is provided, such as the Strategic Plan or Operating Plan.
- Information about the marketplace is provided, as in the report on the size of the OSS market.

8.2. Document History

8.2.1. Version History

Created by TM Forum Publishing Staff

Version Number	Date Modified	Modified by:	Description of changes
v.2.2.0	28/JUN/2015	Alicja Kawecki	
v2.2.1	1/JULY/2015	Alicja Kawecki	Added Admin Appendix page
v3.0.0	11/NOV/2015	Snigdha Mitra	Updates for Fx15.5
v3.0.1	7/DEC/2015	Alicja Kawecki	Formatting and minor cosmetic corrections prior to publishing
v4.0.0	2/JUN/2016	Snigdha Mitra	Updates for Fx16
v4.0.1	6/JUN/2016	Alicja Kawecki	Updated cover; minor formatting/style edits prior to publication for Fx16
v4.0.2	16/AUG/2016	Alicja Kawecki	Updated cover and Notice to reflect TM Forum Approved status
v5.0.0	11/NOV/2016	Snigdha	Updated for Fx 16.5

		Mitra	
v5.0.1	5/DEC/2016	Alicja Kawecki	Updated cover, minor formatting/style edits prior to publication for Fx16.5
V5.0.2	12/JUN/2017	Alicja Kawecki	Updated cover, header, footer and Notice to reflect TM Forum Approved status; applied rebranding

Created by Confluence

Version	Date	Comment
Current Version(v. 2)	**Jun 13, 2017**	**Alicja Kawecki**
v. 1	Dec 05, 2016 11:33	**Alicja Kawecki**

8.2.2. Release History

Release Number	Date Modified	Modified by:	Description of changes
16.0.0	2/JUN/2016	Snigdha Mitra	Updates for Fx16
16.5.0	11/NOV/2016	Snigdha Mitra	Updated for Fx16.5

8.3. Acknowledgments

This document was prepared by the members of the TM Forum Frameworx 16 Big Data Analytics team:

Area	Name	Company	Contact
Use Case Unification	Larry Chesal	Spirent Communications	Larry Chesal
Reference Model	Sophie Nachmann	Orange	Sophie Nachman
Data Governance	Apple Li	Huawei	Juan LI
Data Governance	Abinash Vishwakarma	NetCracker	Abinash Vishwakarma
Maturity Model	Mrinal Moitra	Cognizant	Mrinal Moitra
Business Value Roadmap	Ruchi Banga	Cognizant	Ruchi Banga
ABDR	Gadi Solotorevsky	Amdocs	Gadi Solotorevsky

Other contributors and reviewers are:

- Prachi Sahoo, Ericsson, prachi sahoo
- Ashraf Mohamed, Verizon, Mohamed Ashraf
- Satishkumar Ponnuswamy, Wipro, Satishkumar Ponnuswamy
- Snigdha Mitra, TM Forum, Snigdha Mitra

Original Authors were:

- Wei Dong, Big Data Works, **Co-project leader & Co-author/Editor**
- Dr. Mick Kerrigan, Amdocs Management Limited, **Co-project leader & Co-author/Editor**
- JunPing Wang, Institute of Automation, Chinese Academy, **Co-editor/ Co-author**
- Nikos Tsantanis, Intracom Telecom, **Co-editor/ Co-author**
- Paul Grepps, Teoco, **Co-author**
- Sophie Nachman-Ghnassia, Orange, **Charter co-sponsor and Co-author**

Additional input was provided by the following people:

- Paul Morrissey, Ventraq, **Contributor**
- Steve Cotton, TM Forum, **Contributor**