Bloor

Data Preparation Challenges in the Telecommunications Industry and Voracity from IRI

"

In this paper we will look at the challenges telecommunications companies (telcos) face in ensuring that the data they collect is accessible, secure and of a quality that enables it to be used in a timely and flexible fashion for a diverse array of use cases to improve customer experience, optimise internal processes and meet regulatory requirements. We will discuss the key data preparation tasks and point to the type of solutions that should be considered.

"

Introduction

If we are to believe that organisations are truly data driven, and that their decisions are informed by the data, then telcos should be the best-informed organisations on the planet. The amount of data they capture is truly phenomenal. By 2050, they will be gathering data from 95% of the world's population thanks to mobile devices, wearable tech and wireless information.

In this paper we will look at the challenges telecommunications companies (telcos) face in ensuring that the data they collect is accessible, secure and of a quality that enables it to be used in a timely and flexible fashion for a diverse array of use cases to improve customer experience, optimise internal processes and meet regulatory requirements. We will discuss the key data preparation tasks and point to the type of solutions that should be considered.

In the context of these challenges, we will highlight the Voracity data management platform and how well it tackles the technical challenges of discovering, integrating and migrating data so that it meets all regulatory needs and provides a sound and secure foundation for providing data driven innovation.

The Telco Data Environment

There have been many case studies over the last five years demonstrating the disruption caused by *"data-driven businesses"*, i.e. those using insights to understand customers, automate processes, change their business models and drive new revenues. In the future, this concept will become an integral part of what it takes to compete successfully, allowing organisations to understand and run all parts of their operations, work with their customers and partners, and take part in external activities in new ecosystems. This applies to telecoms operators as much as any other industry.

Telcos handle an immense amount of data from millions of subscribers. A lot of this data comes from voice calls, in the form of Call Data Records (CDR) or the browsing history of websites and apps on their networks. Moreover, telcos also manage location and mobility data at a very granular level since mobile phones are connected to telecom towers. Telcos also handle payments for phone bills and top-ups that identify delinquency and spending behaviour. Finally, many telcos also own digital services such as video streaming, e-wallets or even advertising businesses whose data helps telcos get a complete user profile.

All these data sources should enable telcos to create an extremely granular 360° view of their customers and detailed insights into the running of their organisations that ultimately lead to better customer experiences, more efficient processes and even the development of new business models.

Unfortunately, the volume and variety of data generated by all these data sources is stretching many existing data management infrastructures beyond their limits. For telcos there is an added layer of complexity in that CDRs are typically encoded in a format compliant with a standard called ASN.1 (Abstract Syntax Notation One). The CDRs are not human readable and therefore need an ad-ditional layer of processing to make the data usable for analytics purposes.

Data Management Challenges

As we have already alluded to above, the explosion of data volumes is an obvious, and ongoing challenge. Business managers are aware of this. However, they are less likely to understand the challenges and complexities of getting that data in a state that enables analytics to be able to deliver the business insights they crave.

Time-consuming data preparation tasks

In a recent Anaconda State of Data Science Report, survey responses from 2,360 data professionals, identified that over 45% of data scientists' time was spent on data preparation and cleansing

> **Volume and variety of data generated by all these data sources is stretching many existing data management infrastructures beyond their limits. For telcos there is an added layer of complexity in that CDRs are typically encoded in a format compliant with a standard called ASN.1.**

> **Telcos hold significant volumes of customer information that can only be used for analytics purposes so long as the customers' PII is not revealed. The ability to quickly and comprehensively mask PII so that customer data becomes anonymised is now a critical requirement.**

tasks that take valuable time away from real data science work. So, what is involved?

Data discovery, finding out where all your data is, is the first task. This sounds simple, but over time data from different sources has been stored in separate, siloed databases. Given the amount of data and the variety of sources many organisations don't know, easily, where all their data is. The time and money spent trying to track down all their data sources slows business processes and data in-sights, increases data storage costs, and undermines competitiveness.

Next, the source data has to be profiled to understand its structure, content and interrelations, as well as its quality – which covers its accuracy, completeness, consistency, timeliness and accessibility.

This is followed by extracting the data you want from the primary source, transforming it into a format compatible with the structure of your data warehouse or data lake, and then loading the data to the target environment(s).

Finally, Data Enrichment; this refers to enhancing existing information by supplementing missing or incomplete data with relevant context obtained from additional sources. In other words, it is the process of improving, refining, and augmenting raw data.

Regulatory Compliance and Data Privacy
Thorough data preparation forms an essential grounding for easier compliance with a growing range of data security and data privacy regulations. However, areas like the protection of Personal Identifiable Information (PII) have become increasingly important in recent years.

Telcos hold significant volumes of customer information that can only be used for analytics purposes so long as the customers' PII is not revealed. The ability to quickly and comprehensively mask PII so that customer data becomes anonymised is now a critical requirement.

As a telco, whether you are interested in analytics, in providing self-service data access, in machine learning, cloud migrations, or regulatory compliance, data management in general, and data preparation in particular, data governance is a necessity. Data preparation tasks are labour-intensive, and data quality is often inadequate. Data accuracy can be compromised, data lineage is often unavailable, and data often does not reflect the latest version of data objects in production systems.

The way forward
Historically, these data preparation processes have relied on a stack of tools, often unintegrated and reliant upon different vendors to handle the various data preparation tasks. Now, modern data management platforms offer more comprehensive and integrated solutions. Data Management covers a much wider area than the data preparation tasks we have discussed here, so having a data management platform that has excellent data preparation and governance capabilities that can be acquired and implemented separately might be an important consideration.

Given that many telcos will already have well established analytics, business intelligence, data pipelines, data warehouses, data lakes and other data tools, it is important to look for solutions that provide ease of integration into existing environments.

Implementing a modern data management platform that reduces the manual data preparation workload, improves data quality, masks PII, and simplifies the process of getting data analytics and regulatory compliance ready will promote innovation and help telcos navigate their ever changing business, technology and regulatory environments.

Solution: Voracity from IRI

What is Voracity?

Powered by CoSort (or Hadoop), and built on Eclipse, **IRI Voracity** is a multi-purpose data management platform designed to perform, speed, and consolidate common work in five general areas:

1. Data Discovery – data profiling, classification, search, and metadata redefinition

2. Data Integration – high volume ETL, change data capture, slowly changing dimensions

3. Data Migration – file/data/database type conversion, replication, and federation

4. Data Governance – data quality, PII masking, re-ID risk scoring, test data synthesis

5. Analytics – embedded BI, tie ups to Datadog, KNIME and Splunk, wrangling for the rest

As can be seen in *Figure 1*, Voracity drives solution depth by supporting critical data movement and manipulation requirements – and combining many of these tasks in one job and I/O pass – for a wide variety of data sources and targets.

What Does Voracity Do?

IRI CoSort is the default Voracity data integration engine. Unlike other such products, it is not confined to ETL (extract, transform and load) operations but also performs data replication (change data capture), federation, masking, cleansing, and reporting. Another key point to note about it is that it does not have to transform data in separate steps. You can define jobs that way, but at run time the engine consolidates multiple steps to reduce I/O. Added to the fact that the CoSort run-time engine uses a 2MB, multi-threaded C executable (called SortCL) and loads only the libraries it requires,

Figure 1 –
IRI Solutions
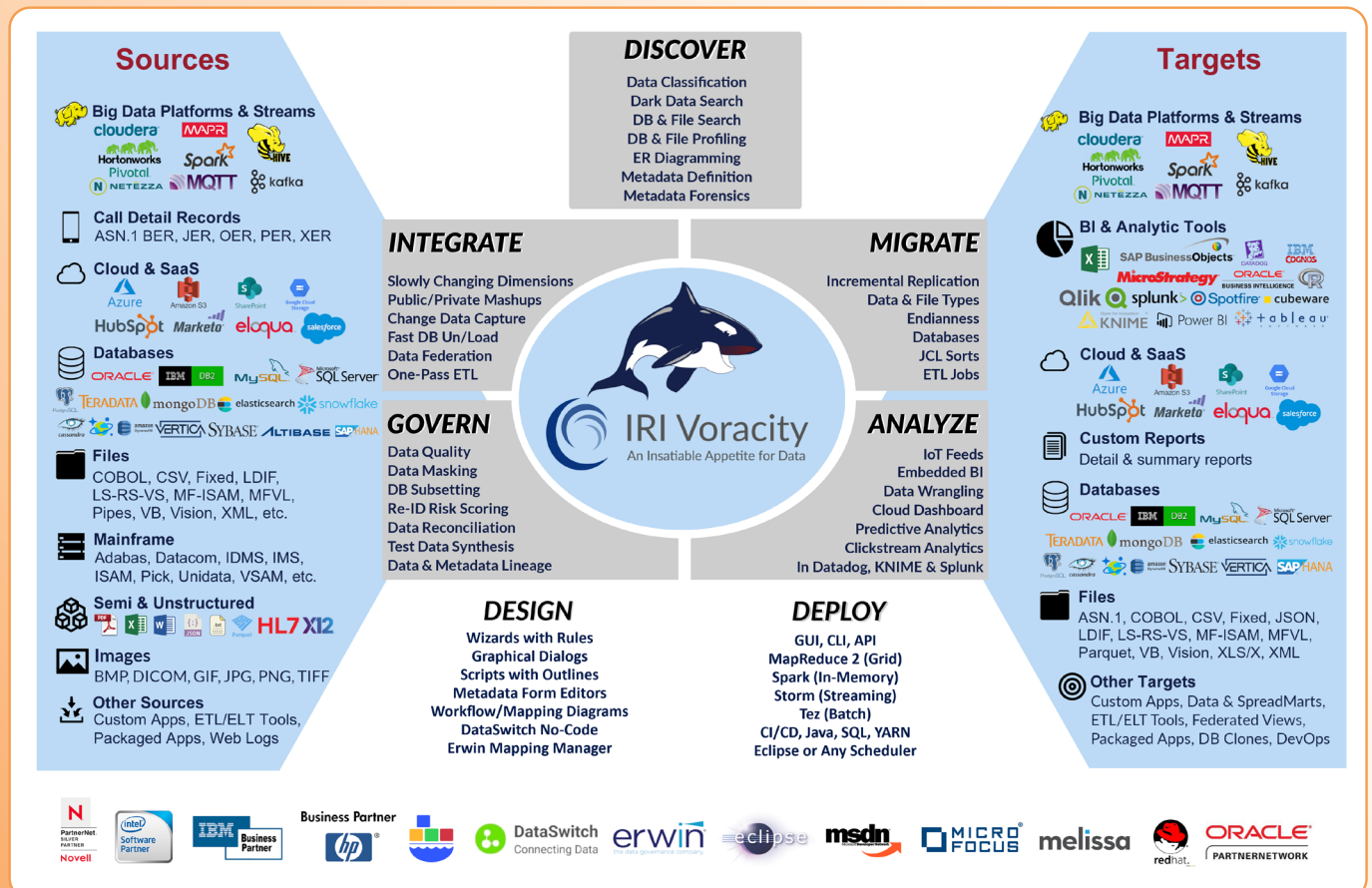
> **Voracity supports the ability to search, parse and protect multiple sources containing semi- and unstructured data with consistent masking functions to support enterprise data integrity.**

and you will appreciate why CoSort has a performance advantage over its competitors.

Note that IRI also offers a Hadoop-based option that does not have the same footprint advantages of CoSort but can often otherwise run in a similar fashion because many jobs developed for native CoSort implementations work without change in Map Reduce 2, Spark, Spark Streams, Storm or Tez. Dataflows are actually stored in files and can be executed from anywhere.

The company offers an extensive range of native connectors (including MQTT and Kafka) plus ODBC support. Not surprisingly given its heritage, it also supports mainframe sources that use COBOL copybooks, EBCDIC and so on. While it does not run on z/OS, it does support mainframe databases as sources and will itself run on z/Linux.

Thanks to ASN.1 support in the SortCL program, users of the Voracity data management platform and its component products can now process CDRs from TAP3 and other encoded data, along with any combination of schema and ASN.1 encoding rules.

While IRI Voracity does not offer a module called *"data quality"*, it does front-end several data quality rules, and provides substantial relevant capability in SortCL jobs, as illustrated in *Figure 2*.

A major strength of IRI Voracity is clearly in its Data Protector Suite. To begin with, IRI has deployed machine learning (including within IRI DarkShield) to support the identification of sensitive data (and we are anticipating that M/L will be implemented more widely across the platform). It also uses natural language processing for this purpose. Once PII is discovered, as mentioned, the company offers significant capabilities when it comes to masking. In particular, dynamic data masking may be proxy-based, run in situ or driven by APIs, and can be mixed and matched with static masking. It is also worth mentioning that Voracity supports the ability to search, parse and protect multiple sources containing semi- and unstructured data with consistent masking functions to support enterprise data integrity.

Finally, given the current predilection for companies to migrate from on-premises data warehouses to cloud-native data warehouses such as Snowflake or Google BigQuery, it is worth noting the availability of IRI FACT, IRI Ripcurrent, and IRI NextForm, which Voracity supports to bolster high volume database migrations using batch or real-time data mapping.
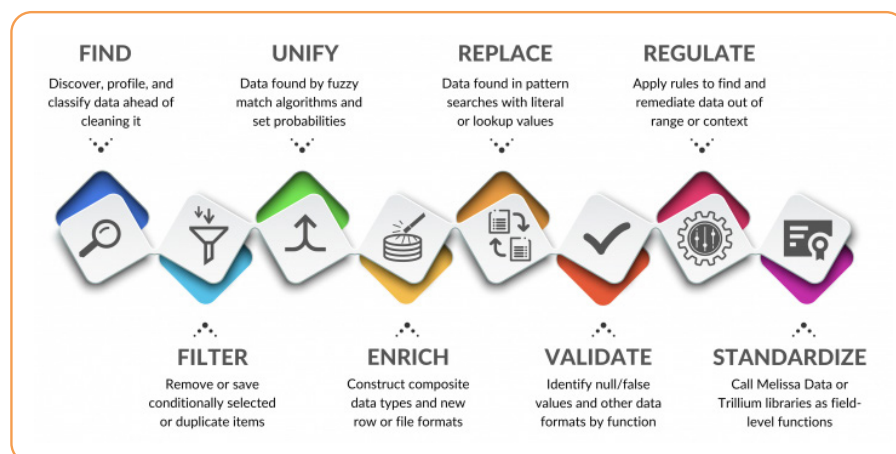
Figure 2 –
Data quality functions
in IRI Voracity

Why should you care?

As a data management platform, one of Voracity's keys strengths is its ability to not just deliver all of the functionality we've described above, but to deliver it within a unified, centralised, integrated platform supported by a common engine. Platform solutions in general are measured by how well they achieve this – after all, without this sort of cohesion they are just a collection of products – and we are pleased to say that IRI does not disappoint. Voracity delivers all of its functionality at speed and at volume, enterprise-wide and at enterprise-scale, which should go a long way in reducing the amount of time spent on data preparation.

That said, it is clear that data integration and transformation lie at the heart of the platform, largely owing to IRI CoSort and its 45-year lineage. What's more, technologies of this type might as well be essential for building data lakes and data warehouses, performing analytics, moving data out of legacy tools or into the cloud, and accomplishing the vast majority of whatever other data manipulation initiatives you can think of. Thence IRI is well-positioned to address all of these various use cases while operating at big data scale and speed.

In addition, functions like cleansing, masking and wrangling can be performed in combination, achieving multiple positive outcomes with only one pass over the data. Moreover, the IRI Workbench acts as something of a central hub, and a common graphical user interface, for actioning almost any particular task within Voracity. Although perhaps it is not the flashiest tool, it is extremely well suited for its intended purpose of providing developers and data engineers with a single point of entry to a wide range of IRI functionality. When seen from this perspective, we would consider it a significant and notable win for developer and data engineer ease of use and productivity.

On the other hand, it is much more advanced when it comes to ETL performance and sensitive data protection than many of its competitors. The company's data migration capabilities will also be a boon in the current environment, as will its relatively attractive price points and licensing options.

> **Functions like cleansing, masking and wrangling can be performed in combination, achieving multiple positive outcomes with only one pass over the data.**

The Bottom Line

The key features of IRI Voracity are the performance that the CoSort engine offers, and the depth of capability it provides in extending its data management platform into the identification and management of sensitive data. If these are important issues for you, then you should seriously consider IRI Voracity. For telcos in particular a wide array of data manipulation and generation functions can be performed with IRI support for ASN.1-encoded data in SortCL-compatible jobs, in ways other solutions for processing data (and much less this type of data) cannot.

The long and the short of it is that Voracity is well-positioned as a holistic platform for data management with particular strengths in data integration and masking.

"

Voracity is well-positioned as a holistic platform for data management with particular strengths in data integration and masking.

"

FURTHER INFORMATION
Further information about this subject is available from
www.bloorresearch.com

About the author
PAUL BEVAN
Navigator, Research Director:
IT Infrastructure

Paul has had a 40-year career in industry that started in logistics with a variety of operational management roles. For the last 33 years he has worked in the IT industry, mostly in sales and marketing, covering everything from mainframes to personal computers, development tools to specific industry applications, IT services and outsourcing. In the last few years he has been a keen commentator and analyst of the data centre and cloud world. Until recently he was also a non-executive director in an NHS Clinical Commissioning Group.

Paul has a deep knowledge and understanding about the IT services market and is particularly interested in the impact of Cloud, Software Defined infrastructure, OpenStack, the Open Compute Project and new data centre models on both business users and IT vendors. His mix of business and IT experience, allied to a passionate belief in customer focus and "grown-up" marketing, has given him a particular capability in understanding and articulating the business benefits of technology. This enables him to advise businesses on the impact and benefits of particular technologies and services, and to help IT vendors position and promote their offerings more effectively.

Bloor overview

Technology is enabling rapid business evolution. The opportunities are immense but if you do not adapt then you will not survive. So in the age of Mutable business Evolution is Essential to your success.

We'll show you the future and help you deliver it.

Bloor brings fresh technological thinking to help you navigate complex business situations, converting challenges into new opportunities for real growth, profitability and impact.

We provide actionable strategic insight through our innovative independent technology research, advisory and consulting services. We assist companies throughout their transformation journeys to stay relevant, bringing fresh thinking to complex business situations and turning challenges into new opportunities for real growth and profitability.

For over 25 years, Bloor has assisted companies to intelligently evolve: by embracing technology to adjust their strategies and achieve the best possible outcomes. At Bloor, we will help you challenge assumptions to consistently improve and succeed.

Copyright and disclaimer

This document is copyright **©Bloor 2023**. No part of this publication may be reproduced by any method whatsoever without the prior consent of Bloor Research. Due to the nature of this material, numerous hardware and software products have been mentioned by name. In the majority, if not all, of the cases, these product names are claimed as trademarks by the companies that manufacture the products. It is not Bloor Research's intent to claim these names or trademarks as our own. Likewise, company logos, graphics or screen shots have been reproduced with the consent of the owner and are subject to that owner's copyright.

Whilst every care has been taken in the preparation of this document to ensure that the information is correct, the publishers cannot accept responsibility for any errors or omissions.