Railway Communication Network Digital Reporting System: A Case Study

Alexander Suleykin

Peter Panfilov

Follow this and additional works at: https://aisel.aisnet.org/menacis2021

This material is brought to you by the MENA at AIS Electronic Library (AISeL). It has been accepted for inclusion in MENACIS2021 by an authorized administrator of AIS Electronic Library (AISeL). For more information, please contact elibrary@aisnet.org.
Railway Communication Network Digital Reporting System: A Case Study

Completed Research Paper

Alexander Suleykin\textsuperscript{a} and Peter Panfilov\textsuperscript{b}

\textsuperscript{a} Russian Academy of Sciences’ V. A. Trapeznikov Institute of Control Sciences, Profsoyuznaya St. 65, Moscow 117342, Russian Federation

\textsuperscript{b} National Research University – Higher School of Economics, Myasnitskaya St. 20, Moscow 101000, Russian Federation

Abstract

The digital transformation of the railway communication network operations is contributing to an increase in data volumes across various functional areas, collecting and processing information from equipment, customers, and the company's various business process management systems. The implementation of a specialized solution for analysis and operational monitoring was a logical extension of the development of data handling, opening up opportunities for building both operational reporting and predictive analytics, distribution and unification of reporting forms, which the reporting in production systems and manual analytics generation tools used before did not allow to achieve. In this paper, we propose to use a digital framework based on open-source big data technologies with different distributed services to process and store KPIs data from Railway Communication Network. We also review literature about challenges and solutions of Data Lake in KPI’s Railway Communications Network domain. To assess the applicability of the framework proposed, incidents and event data relating to the production business processes have been tested on the actual railway communication network. The suggested framework effectively processed the data and the implementation setup settings were presented, as well as the results of experiments.

Keywords: Digital Ecosystems; Railway KPIs; Digital Reporting; Distributed Systems; Open-Source Technologies; Big Data.

Introduction

Digitalization has become a priority at the micro and macro levels. Software is rightly seen as a tool that can make a real difference. Today, to remain competitive in the global market public corporations and companies from different sectors of the economy have to leverage their competencies in ICT taking advantage of new opportunities that digital transformation brings for new business models and demands for innovations in their respective ecosystems.

It is critical that the system be able to recover and tolerate faults without failing, which can be handled by hardware, software, or a combined solution leveraging load balancers for reliability, which is focused on a continuous service without any interruptions, availability, which is concerned with the system’s reading readiness, and security, which prevents any unauthorized access.

The digital transformation of the railway communication network operations is contributing to an increase in data volumes across various functional areas, collecting and processing information from equipment, from customers, and from the company’s various business process management systems. The implementation of a specialized solution for analysis and operational monitoring was a logical continuation of the development of data handling, opening up opportunities for building both operational reporting and predictive analytics, distribution and unification of reporting forms, which the reporting in production systems and manual analytics generation tools used before did not allow to achieve.

A large railway company in Russia plays an important role in the development of the transportation and logistics industries by providing communications railways, automate operational and analytical reporting to elevate business analytics and data management to a new level.
Digital Reporting plays an important role in the whole digital transformation. Reports, based on the digital services, are the core improvement driver for analytics of KPIs data and formulation of new Data-driven solutions made by decision makers. In order to create reports, a lot of manipulations with data to be done: integrated many different data sources, required data transformations, ETL-processes execution, data pipeline orchestration, normalizations and filtering, creation of dashboards and graphs in BI system and finally visualization.

In this paper, we propose to use a digital framework based on open-source big data technologies with different distributed services to process and store KPIs data from Railway Communication Network. We also review literature about challenges and solutions of Data Lake in KPI’s Railway Communications Network domain. To assess the applicability of the framework proposed, incidents and event data relating to the production business processes have been tested on the real railway communications network. The suggested framework effectively processed the data, and the implementation setup settings, as well as the implementation results, were given.

The rest of this paper is organized as followed: first, we review Data Lake for Railway Communication Network Reporting. Second, we present the main data processing challenges for Railway communication network reporting processes and main problematic issues. Third, we provide an architectural overview of the proposed solution and its main Services. Fourth, we design the functional architecture of the Digital Communication Network Reporting Framework for Railway Reporting and describe the main Services and data processing pipeline in our practical implementation of the framework. After that, experimental setup parameters (software and hardware), which were used in our test implementation are described. Finally, the Communication Network Digital Reporting pipeline is modeled and implemented based on described data and proposed architecture.

**Data Lake Technology for Digital Reporting**

Data Lake (DL) is a Big Data analysis solution which ingests raw data in their native format and allows users to process these data upon usage. Data ingestion is not a simple copy and paste of data, it is a complicated and important phase to ensure that ingested data are findable, accessible, and interoperable and reusable at all times (Zhao et al., 2021).

Over the past two decades, we have witnessed an exponential increase of data production in the world. So-called big data generally come from transactional systems, and even more so from the Internet of Things and social media. They are mainly characterized by volume, velocity, variety and veracity issues. Big data-related issues strongly challenge traditional data management and analysis systems. The concept of Data Lake was introduced to address them. A data lake is a large, raw data repository that stores and manages all company data bearing any format. However, the data lake concept remains ambiguous or fuzzy for many researchers and practitioners, who often confuse it with the Hadoop technology. Thus, we provide in this paper a comprehensive state of the art of the different approaches to data lake design. We particularly focus on data lake architectures and metadata management, which are key issues in successful data lakes. We also discuss the pros and cons of data lakes and their design alternatives (Sawadogo and Darmont, 2021).

Although big data has been discussed for some years, it still has many research challenges, especially the variety of data. It poses a huge difficulty to efficiently integrate, access, and query the large volume of diverse data in information silos with the traditional `schema-on-write` approaches such as data warehouses.

Data lakes have been proposed as a solution to this problem. They are repositories storing raw data in its original formats and providing a common access interface (Hai et al., 2021).

Data lakes are today widely being used to manage the vast amounts of heterogeneous data sources in enterprises. Different from classical data warehouses, the idea of data lakes is that data does not need to be organized and cleaned upfront when data is loaded into the warehouse. Instead, data lakes follow a more “lazy” approach that allows enterprises to store any available data in its raw form. This raw data is organized and cleaned once it is needed for a downstream task such as data mining or building machine learning models. However, due to the sheer size of data in data lakes and the absence (or incompleteness) of a comprehensive schema, data discovery in a data lake has become an important problem (Langenecker et al., 2021).
The most successful work about data lake architecture, components and positioning is presented, because the emphasis is on data governance and more specifically on the metadata catalog. Initially regarded as low-cost storage environments, data lakes are now considered by companies as strategic tools due to their potential ability to give data a high value (Laurent et al., 2020).

By moving data into a centralized, scalable storage location inside an organization – the data lake – companies and other institutions aim to discover new information and to generate value from the data. The data lake can help to overcome organizational boundaries and system complexity. However, to generate value from the data, additional techniques, tools, and processes need to be established which help to overcome data integration and other challenges around this approach (Mathis, 2017).

The DLAF consists of nine data lake aspects to be considered, their interdependencies, and methodology to choose appropriate concepts for each aspect. The evaluation showed that the DLAF can be applied in two ways: 1) it can be used to identify missing aspects in existing data lake implementations and provide pointers towards re-design of the architecture. 2) The DLAF can be used to define a novel comprehensive data lake architecture (Giebler et al., 2021).

As known, railways networks are very complex systems formed by many components and devices which must be properly coordinated, continuously monitored, and timely operated, in order to ensure the correct and efficient functioning of the traffic over the whole railway network, while satisfying the highest standards of safety for the fleet of operating vehicles, which is a very critical aspect in the railroad industry (Muniandi, 2021). In this work, we consider different approaches to railway network reporting problem. The project based on open source big data distributed, scalable and fault tolerant data processing pipeline technologies to solve all problems and issues related to railway communication network reports.

The occurrence of the concept of a data lake to meet such big data problems is enlightening and will most likely be considered in any relevant big data strategy. This idea is still on the way to prove itself out and inevitably it gives rise to much attention as well as much criticism. Luckily, more and more positive voices towards data lakes are emerging and give highly appreciation to the concept and even propose some workable and innovative suggestions to make improvement to the practical implementation (Singh and Ahmad, 2021).

**Data Lake Modern Design Patterns, Methodology and Approach**

The current Data Lake design tendency has been shifted to hybrid architectures, where different architectural patterns are met – both cloud (or private cloud) and bare metal, as well as different system types for data storage – from MapReduce Hadoop systems with relatedly slow MapReduce engine, to classical MPP-systems and In-Memory data storages. Each system in different storage types is oriented on optimal storage, read and write speed, and CPU usage.

According to the requirements of the company, different limitations and preferred technology stack, there many Data Lake design patterns are widely used nowadays – bare metal, public and private cloud, MPP-systems both Open-Source and not, MapReduce Hadoop and Hadoop ecosystems, In-Memory Data Grids and Quick Marts storage systems.

There exit many Data Lake building methodologies in terms of data models and transformations. The methodologies research was out of the scope of this paper. In out experiment, Kimball-based Data Lake design was used. In the proposed prototype system, we have used virtualized private cloud and MapReduce Hadoop for main Data Lake storage system and engine, and PostgreSQL for data marts calculations and data storage.

In terms of the existing approaches for Data Lake design, we proposed to use hybrid architecture – Apache Hadoop for historical storage, and PostgreSQL for data marts, and all digital services are developed using the Open-Source technologies. This approach was applied to the Railway Communication Network Reporting domain, which is quite new and not presented in the current state of the art.
Challenges of Communication Network Digital Reporting

Data Integration and Orchestration

Reporting is, in reality, one of the most extensive business processes in any company. It implies a wide range of data integration initiatives involving a wide range of data sources, formats, protocols, processes, and tools. Reporting necessitates that data be ready and available by a certain date, which necessitates the organization and coordination of numerous humanitarian efforts in order to achieve such a pipeline on time. Furthermore, many ETL steps run in a specific order and follow or are followed by other operations. Because we need to build an effective data pipeline with suitable phases, integrate many different data sources, and adhere to the Reporting SLAs, the issue of a data integration methodology is quite important. With the support of a central service, i.e. an Orchestration Service, and ETL Services, which launch specific ETL jobs, the Communication Network Digital Reporting Framework meets all of these issues of data integration and ETL-process orchestration.

Business Intelligence and DWH Interaction

The purpose of a business intelligence system is to visualize reporting data. Data Lake or Data Warehouse are the primary data sources for BI systems, and data loading in BI systems typically begins once all data transformations in Data Lake have been completed successfully. Because preparing data by exact time is not always achievable in real-world situations, this event-driven strategy has proven to be beneficial. The amount of data collected for reporting from source systems fluctuates from day to day, and data may be ready for BI processing at different times on different days. As a result, the BI service and Storage Service must communicate in such a way that the BI system can only begin loading data when a particular signal from Storage Service through API is generated. A specific BI service is built in our Communication Network Digital Reporting Framework implementation that launches BI jobs once all data marts are ready.

Architecting the Railway Communication Network Digital Reporting

A System for Communication Network Digital Reporting

Communication network reporting system is a highly important system in any organization. In order to build a robust, reliable, fault-tolerant, scalable, and high-loaded data pipelines for communication network reporting that follows strict SLAs it is proposed to use the following architecture of Reporting that conforms to Distributed scalable and fault tolerant system concept (Figure 1).

The Communication Network Digital Reporting Framework Includes:

1. **Orchestration Main Service** – The services responsible for orchestrating all other services: ETL, Data Lake. It is the core part of the system, and all other integrations and new service designs will also be orchestrated on this service. It has a flexible distributed worker assignment algorithm, and only available, idle, and active worker nodes can assign tasks. The solution provides a robust, fault-tolerant, reliable, and scalable design for the ecosystem;

2. **External Integration Service** - A service that communicates with BI service. The role of this service is to activate the BI service, and the loading of BI data should start immediately. In addition, the role of this service is also the interaction with any external ecosystem, and it may be integrated into the reporting ecosystem of the digital communication network in the future;

3. **Core ETL Services** - ETL data transformation and processing services across all data pipelines, right through to visualization. ETL can appear in any type of job, but preferably they should be distributed and fault-tolerant;

4. **Data Lake** - A service that stores all data in a distributed and reliable way such as historical Data, Detailed Data, and Data Marts. All data are stored in this service and all results of the processing are also stored in this service. Data integrate with all source systems, data visualization for reporting. On the basis of visualized periodic reports, the administration will make decisions based on data, which will inevitably lead to the creation of new services and new ecosystems.
Communication Network Digital Reporting System for Railway Sector

Based on the Communication Network Reporting Framework concept described above, we have proposed Railway Communication Network Reporting System Architecture for Railway Communication Network Sector that includes specific ETL-processes for Railway reporting. Figure 2 depicts the functional architecture of the proposed Digital Reporting System for the Railway Communication Network system.
The Railway Communication Network Digital Reporting Framework is built using open-source Big Data technologies, which satisfy following requirements:

- Horizontally scalable Services for data storage – a layer for data loads to the archival and distributed storage.
- Service with distributed data processing engine capable of processing large volumes of data.
- Fault-tolerant Service – the solution works without incurring data loss in case of a single failure.
- All Services should be orchestrated and combined in one main Service that can run other Services.
- Service with a relational database for data access to different BI tools.
- Be Open Source.

On the basis of these criteria, we have defined the following Open-Source Big Data technologies for the implementation of the proposed Railway Communication Network Digital Reporting system (Table 1):

<table>
<thead>
<tr>
<th>No.</th>
<th>Component</th>
<th>Definition and role in System</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Spark</td>
<td>Distributed in-memory System for high-load data processing (spark.apache.org). Spark has been used to create all main ETL Services, which processed data and move it through all Services</td>
</tr>
<tr>
<td>2</td>
<td>HDFS</td>
<td>Distributed fault tolerant file system optimized for storage for processing large volumes of data (hadoop.apache.org). A main Service for distributed data storage of Railway data for Reporting in test implementation</td>
</tr>
<tr>
<td>3</td>
<td>PostgreSQL</td>
<td>Relational database for universal data access from BI systems (<a href="http://www.postgresql.org">www.postgresql.org</a>). In test experiment, we used PostgreSQL as a main service for Serving layer and access to BI service</td>
</tr>
<tr>
<td>4</td>
<td>Airflow</td>
<td>Orchestrator and Scheduler (airflow.apache.org). In our experiment, Airflow was main service for other Services orchestration and scheduling</td>
</tr>
</tbody>
</table>

Table 1. Selected Technologies for Railway Communication Network Reporting System Modeling and Implementation

**Main ETL and Storage Services Dataflow Description**

Data can be traced from various consumer sources. Managing data is one of the most serious challenges faced by organizations today. Organizations are adopting the data lake models because lakes provide raw data that users can use for data experimentation and advanced analytics. A data lake could be a merging point of new and historic data, thereby drawing correlations across all data using advanced analytics. A data lake can support the self-service data practices. This can tap undiscovered business value from various new as well as existing data sources (Singh and Ahmad, 2021).

In our research, we have built a powerful multi-staged data processing pipeline that combines two major super stages or layers – Data Storage Layer and Serving Layer (Suleykin and Panfilov, 2020). Here Storage Layer and Serving Layer have their own Layers (sublayers), which are used for methodological correctness of data load. The data processing pipeline of the whole data movement is strict and should go through the following sublayers/stages inside Serving and Storage Layers as it is depicted in Figure 3.

The details of the Storage Service implementation such as open-source technology used, functionality, and data transformations are presented in Table 2.
<table>
<thead>
<tr>
<th>Name</th>
<th>Abbreviation</th>
<th>Location</th>
<th>Definition and functions</th>
<th>Transformatios</th>
</tr>
</thead>
<tbody>
<tr>
<td>Staging Buffer Area</td>
<td>STG/BUF</td>
<td>HDFS</td>
<td>The area where temporary data is accumulated in the source’s format without any changes.</td>
<td>No</td>
</tr>
<tr>
<td>Staging Exchange Area</td>
<td>STG/EXCH</td>
<td>HDFS</td>
<td>The region in between that to be used to create the following ETL processing packet. All accumulated data is relocated from the buffer. A unique BATCH ID is assigned to it.</td>
<td>BATCH_ID</td>
</tr>
<tr>
<td>Staging Archive Zone</td>
<td>STG/ARCH</td>
<td>HDFS</td>
<td>Storage of the entire archive of incoming messages without any storage format change. Following successful processing, incoming messages are archived.</td>
<td>Archiving and enlarging storage files</td>
</tr>
<tr>
<td>Operational Data Store</td>
<td>ODS/FULL</td>
<td>HDFS</td>
<td>The original data scheme is stored in this section, however, it is simplified to a single binary form of storage. It has a complete history of all changes and deletions.</td>
<td>Conversion to binary storage format</td>
</tr>
<tr>
<td>Batch View</td>
<td>ODS/BV</td>
<td>HDFS</td>
<td>It stores a snapshot of the current state of objects, with no change history or deleted data.</td>
<td>Calculation of the actual data slice</td>
</tr>
<tr>
<td>Detail Data Store Staging</td>
<td>DDS_STG</td>
<td>Postgre SQL</td>
<td>Layering by batch for each source system, a unique instance is produced. Only between downloads is one-to-one data transferred from HDP and stored. It is possible to get a full data load as well as just line modifications (deltas).</td>
<td>Conversion from object to relational storage</td>
</tr>
<tr>
<td>Detail Data Store</td>
<td>DDS</td>
<td>Postgre SQL</td>
<td>The current data slice’s layer is displayed in a relational format. Putting together a single data model (without unification). Changing the keys (generation of internal storage IDs).</td>
<td>All data access to DDS should be organized using views on top of DDS layer itself</td>
</tr>
<tr>
<td>Detail Data Store View</td>
<td>DDS_v</td>
<td>Postgre SQL</td>
<td>The layer for data access to DDS</td>
<td>Data unification, denormalization, and aggregation</td>
</tr>
<tr>
<td>Data Mart</td>
<td>DM</td>
<td>Postgre SQL</td>
<td>The showcases of a group are organized by a specific attribute, most commonly the subject area. Contains a unified set of detailed information.</td>
<td>Calculation of derived indicators for a specific report</td>
</tr>
<tr>
<td>Data Mart Reporting</td>
<td>DM_rep</td>
<td>Postgre SQL</td>
<td>The last tier of reporting. Data from it is exclusively utilized to present in BI tools. It is not permitted to construct some reports based on the findings of others. Only when the information in the DM layer is transferred.</td>
<td></td>
</tr>
</tbody>
</table>

Table 2. Storage Service description of data processing pipeline
Experiments and Results

The proposed Digital Communication Network Ecosystem has been tested using Reporting KPIs data from one of the largest Railway company in Russia. Data size is 10 Gb for Incidents and 3 Gb for Events. The data consists of two tables of incidents and events. Incidents have 7 entities and 13 references. Events have 1 entity. And all of those are used in our experiments for testing data processing workflow implementation. Entities for incidents were fault_data, fault_data_frends, Lrp_responsible, Lrp_plan_work, Lrp_approvers, Lrp_method_work, Lrp_transfer_work, references were lrp_status, Lrp_type_lrp, Lrp_type_problem, Xconseq_tree, Xmotive_tree, Snetsystem, Node, Sstation, Sorganizationtype, Sorganizationlevel, Sorganization, Sec_user, Lrp_code_close. For events a fault_all entity is used.

Experimental Setup Parameters

Tests of the proposed Railway Communication Network Digital Reporting Framework were carried out in the simulation testbed. Platform servers are located on virtual machines based on the same physical server. The testbed was deployed using Microsoft Hyper-V virtualization based on dedicated servers, and test runs were conducted using virtual machines. The configuration of test virtual machines is shown in Table 3:

<table>
<thead>
<tr>
<th>VM host</th>
<th>CPU (vCores)</th>
<th>RAM (Gb)</th>
<th>HDD (Tb)</th>
<th>Services and their roles</th>
</tr>
</thead>
<tbody>
<tr>
<td>pg-source_system.bd</td>
<td>8</td>
<td>16</td>
<td>1</td>
<td>Postgre SQL server – data base for Serving Layer</td>
</tr>
<tr>
<td>bd-hdp-edge01</td>
<td>8</td>
<td>32</td>
<td>1</td>
<td>Airflow server – ETL orchestration</td>
</tr>
<tr>
<td>demo-hdp-01</td>
<td>4</td>
<td>32</td>
<td>1</td>
<td>PostgreSQL metadata; Zookeeper</td>
</tr>
<tr>
<td>demo-hdp-02</td>
<td>4</td>
<td>32</td>
<td>1</td>
<td>Spark2; YARN; Zookeeper</td>
</tr>
<tr>
<td>demo-hdp-03</td>
<td>4</td>
<td>16</td>
<td>1</td>
<td>HDFS Name Node; MapReduce2; Hive; YARN; Zookeeper</td>
</tr>
<tr>
<td>demo-hdp-04</td>
<td>4</td>
<td>16</td>
<td>1</td>
<td>HDFS JournalNode; Zookeeper Server; Spark2 Thrift Server</td>
</tr>
<tr>
<td>demo-hdp-05</td>
<td>4</td>
<td>16</td>
<td>1</td>
<td>HDFS Datanode; YARN Nodemanager</td>
</tr>
<tr>
<td>demo-hdp-06</td>
<td>4</td>
<td>16</td>
<td>1</td>
<td>HDFS Datanode; YARN Nodemanager</td>
</tr>
<tr>
<td>demo-hdp-07</td>
<td>4</td>
<td>16</td>
<td>1</td>
<td>HDFS Datanode; YARN Nodemanager</td>
</tr>
<tr>
<td>demo-hdp-08</td>
<td>4</td>
<td>16</td>
<td>1</td>
<td>HDFS Datanode; YARN Nodemanager</td>
</tr>
<tr>
<td>demo-hdp-09</td>
<td>4</td>
<td>16</td>
<td>1</td>
<td>HDFS Datanode; YARN Nodemanager</td>
</tr>
<tr>
<td>demo-hdp-10</td>
<td>4</td>
<td>16</td>
<td>1</td>
<td>HDFS Datanode; YARN Nodemanager</td>
</tr>
</tbody>
</table>

Table 3. Testbed configuration
**Test Results**

As a result of modeling data processing using Apache Airflow, the Directed Acyclic Graph (DAG) is created that shows Communication Network Digital Reporting Framework Service in action. Thus, after successful execution of DAG in Airflow, the DAG itself changes to a green color indicating that all DAG steps have been finished properly as show in figure 4 and figure 5 for incidents (Part 1 and 2 as the DAG is too big to screenshot in one figure), and figure 6 and figure 7 for events:

**Figure 4. Incidents Part 1**

**Figure 5. Incidents Part 2**
Figure 6. Events Part 1

Figure 7. Events Part 2
Conclusion

In our paper, a Railway Communication Network Digital Reporting Framework for efficient KPIs data processing using Big Data technologies has been created as well as methodology for operational and analytical KPIs data preparation.

We have reviewed literature for data processing and storage for Railway Communication Network sector and discovered that there are no such frameworks and architectures for digital reporting, which are open-source, distributed, scalable and fault-tolerant. Additionally, we regarded data processing challenges for the Railway Reporting sector such as Integration, orchestration and BI subsystem interaction issues. Digital Ecosystem-based Communication Network Reporting Framework has been constructed with the main concept of distributed services and Open-Source Big Data technologies. We propose to use our Open-Source Communication Network Digital Reporting Framework for building distributed, fault-tolerant, scalable, and high-loaded data pipelines for Railway Communication Network Digital Reporting.

The proposed digital reporting framework has been tested on the basis of the actual Railway Communication Network incidents and events data, and successful implementation of data processing with all services such as ETL-Agents, Storage, Integration and BI interaction have been performed. The implementation of the Railway Communication Network Digital Reporting system and proposed KPIs methodology in test environment has demonstrated the architectural framework's versatility and its possible application to other domains and use cases.

Acknowledgements

The reported study was partially supported by RFBR grant No 20-07-00958 and the HSE University Project Group Competition in 2020 — 2022.

References