Validating Enterprise Data Lake Using Open Source Data Validator – An Airline Industry Case Study

Reach the community at users@collaborate.jumbune.org

Download Jumbune from http://www.jumbune.org

An Open Source initiative LGPLv3 licensed

Table of Contents

Overview

A renowned Trans Pacific Airline, ranks among the top international airline company in terms of number of passengers carried. The company operates two of the world's longest non-stop flights and on an average 58 flights, between major cities. By 2013, the airline expanded the fleet to six Airlines A310s and eight Airlines A500s and builds the network to cover 30 destinations across the world.

Business Challenge and Future Proofing

The Airline maintained its information system in relational data store. In order to suffice specific analytical and future business needs, the airline company consolidated data sources within various departments such as Human Resources, Operations, Sales, Maintenance, Customer Relations, Safety, Logistics and Revenue Accounting. All the data from the passenger itinerary and boarding details, maintenance logs, cargo tracking, fuel load, ticket prices, concession and seating, crew details is added onto the data hub. The organization wanted to mine all its data (with large volume, variety and velocity) efficiently and effectively, traditional databases were inefficient perform basic operations and analytics across organizational silos.

VP, Information Technology, recommended the creation of a data lake to consolidate and store the data in a single repository that will solve to all current and future analytical needs. The airlines raw data includes:

1. Operation department includes information related with aircraft such as aircraft details, flying details, crew details, catering details etc.
2. Customer Relations department includes information related with passengers such as personal information, aircraft details, cabin details - First class, Business class, Economy class, etc.
3. Sales department include information related with manual and online bookings such as Passenger information, Booking information, Ticket details, etc.
4. Cargo details……. (Add tracking, weight, etc.)
5. Flight maintenance records, spare part records, safety checking,
6. Ticket pricing (across cities, discounts etc.)
7. Human Resources department includes information of employees such as their personal information, salary, attendance details, designation, etc. ,

Big Data is more about processing large volumes of data. Hadoop, being scalable, reliable and economical, was undoubtedly the preferred choice for storing and analyzing batch data.

Hadoop provides the ability to store large scale enterprise data on Hadoop Distributed File System (HDFS) and analyzing this huge data using execution engines such as MapReduce, Hive, Storm, etc. HDFS is a Java-based file system that provides scalable and reliable data storage that is designed to span large clusters of commodity servers. HDFS is highly fault-tolerant and is designed to be deployed on low-cost hardware. HDFS provides high throughput access to application data and is suitable for applications that have large data sets.

Finding Anomalies using Jumbune's Data Validation

The Engineering team initiated installations and configurations. They created a test environment of 5 nodes Apache Yarn cluster and the preliminary Hive based analytics worked as per the expectations.

Moving to production, they configured a Hadoop commercial distribution of 100 node cluster. After a month, the management noticed that the analytical reports generated by the jobs were erroneous. One of the erroneous use case was generating an analytical report consisting of information on passenger's carry-on luggage with respect to the age, gender of the passenger and also the ratio of the carry-on with the checked-in luggage.

The engineering team spent lot of man hours in writing a customized MapReduce program to uncover the root cause in the analytical logic and later they figured out that the actual problem was caused by the inconsistent data ingested by a malfunctioning ETL instance operating from one of the airports in the new route. The updated policy of the Airport Authority refrained the airlines from recording the actual weight of the carry-on luggage. This introduced number of anomalies into the data hub that led to the erroneous analytics.

Jumbune Data Validation MapReduce job analyze batch, incremental data files kept on HDFS and provides generic categories of validations: Null, Regex and Data Type. Jumbune gives feasibility to analyze TB's of data in comparatively less time and also helps in finding anomalies. The engineering team ran Jumbune's Data Validation module. Jumbune has its own customized MapReduce data validation framework that generically validates data on HDFS. Jumbune is highly optimized, can be operated remotely, user friendly. Only the HDFS path and validations on the fields needs to be provided as the input to the data validation module.

The engineering team ran Jumbune's Data Validation module with null check on all the fields and found that carryon luggage field contains null values. Jumbune analyzed HDFS data and presented the analytics data of number of null values in the data. Furthermore, they found three more use cases where Jumbune's Data Validation module was beneficial to them for finding data anomalies. The use cases were:

1. In order to suffice marketing needs the team wanted to check how many passengers did not enter their mobile number.
Solution: Marketing team applied a null check on mobile number field and ran Jumbune. Jumbune launches its MapReduce program, analyzed HDFS data and presented the analytics report which listed number of passengers who didn't enter their telephone numbers.

2. The management team, required to take customer feedbacks for their flight experience. They found that the most feasible way is to send SMS to all the passengers. They did not know whether the data type validation was applied on phone number field or not.
Solution: With the help of engineering team they launched a Jumbune Data Validation job with data type check on phone number field. The analytics report listed 102 passengers out of millions of entries who entered wrong phone numbers.

3. Airlines sales manager observed a fall in their sales. To know the reason, marketing team created an email survey for the passengers which aimed to know the number of passengers filling incorrect email ids.
Solution: They required knowing the number of passengers who gave improper email ids. With the help of engineering team they launched a Jumbune Data Validation job with regular expression check on email field and the analytics report listed 136 passengers who entered wrong email ids.

Result: Analytical Anomalies report

Analyzing enterprise data results in significant loss in revenue and time, Jumbune's Data Validation has helped this organisation to get analytical report of anomalies in data hub..

The Airlines engineering team downloaded Jumbune from http://jumbune.org/