DATA QUALITY MANAGEMENT ON HADOOP

Hadoop Data Lake

A Data lake is a pool for storing non-relational data as-is, like Sales & Transactional Data, Log Files, Streaming Data, Customer surveys, Social networking comments into a single environment for processing and analyzing to derive future business decisions. Since data is created at a fast velocity and with a large variety, Hadoop data lake is used frequently as its less expensive since it uses commodity hardware and Hadoop being open source. Companies like Banking, Retails, Medical, Aeronautics etc. are using this consolidated data for various analytics, like Customer Journey Analytics, Customer 360, Historical Analytics, Predictive, Prescriptive to even Deep Learning based Analytics and many other ad-hoc or research analytics.

Key elements of an Enterprise Hadoop Data Lake

Following are few key element of data lake

Figure 1

1. **Data Quality** – A measure to evaluate the correctness of collected (or acquired) big data sets.
2. **Data Profiling**– Statistical assessment of data values within a data set for consistency, uniqueness and logic as per the business rules.
3. **Data Governance** – Overall management of the availability, usability, integrity, and security of the data.
4. **Data Lineage** – A metadata defining from where the data came from and how was it calculated.

Consequence of Poor Data Quality

Recent studies have shown that quality of aggregated data in data lake is far from perfect. Poor quality data can bring tangible loss to company in terms of loss of revenue, inefficient service operations, Higher Cost of Maintenance. That's why Data validation is a critical process to recognize and improve data quality.

Jumbune

Identify & Report data inaccuracies

Data Cleansing functionality to correct identified inaccuracies

Allow organizations to maintain a higher level of data integrity

Meter Quality of Data over Ingestions

Gauge Data Insights

http://jumbune.com

contact-at-jumbune.com

+1 408 252 7111

Impetus Technologies, Inc.

720 University Avenue, Suite 130
Los Gatos, CA 95032, USA

Big Data Quality Validation Process

Data validation process is to ensure data validity, data completeness and data consistency and validate data is trustworthy, accurate and meaningful. It's been accessed that more than half of the time spent in big data projects goes towards data cleansing and preparation. This section discusses the validation process for big data. As shown in Fig.2, data collection, data cleaning, data transformation, data loading and results report are the necessary data validation process.

The detailed illustration for data validation process is as follow

1. **Data Collection: -** In this stage, the data is collected from several types of data sources, data marts and data warehouses, such as from Emails, Database, CSV files, excels, File System etc.

2. **Data Lake Loading:** - In this stage, data loading activities in which data are loaded into a big data repository, for example, a Hadoop environment (HDFS) or a NoSQL Big database. Depending upon the requirement of the customer updating extracted data is frequently done on a daily, weekly, or monthly basis

3. **Data Validation: -** In this stage the data is checked data quality, for example, check the data type, formats, phone numbers, numerical values, positive/negative values, greater than previous, greater than range, less than range, lesser than next/previous, etc.

4. **Data Cleansing:-**In this stage, the data is cleansed by correcting or removing the corrupted or inaccurate records from a record set, table or database. The major purpose is to detect and identify incomplete, incorrect, inaccurate, irrelevant data parts in data sets.

5. **Data Transformation: -** In this stage, set of data values from the data format of a source data system into the data format of a destination data system.

6. **Data Analysis and Reporting: -** The last step is to write a document to report data validation results. Data consistency, data completeness and other data quality dimensions are completely analyzed with goal of discovering useful information.

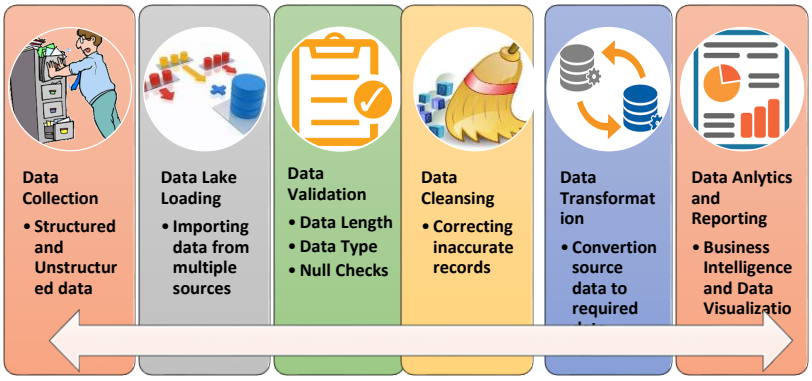

Figure 2

Jumbune – Data Quality Management

Jumbune impeccable Analyze Data module conform to Data Quality Validation Process. Jumbune Analyze Data module have very comprehensive feature like Data Validation, Data Profiling, Data Quality TimeLine, Data Cleansing as shown in Fig.3. Jumbune architecture of executing self-contained MapReduce job analyses batch and incremental data files kept on HDFS. Jumbune gives feasibility to analyze TB's of data in comparatively less time and also helps in finding anomalies using generic categories of validations: Null, Regex and Data Type. Jumbune Analyze Data module not only creates reports which depicts HDFS data discrepancies it also validates HDFS data against data violation constraints.

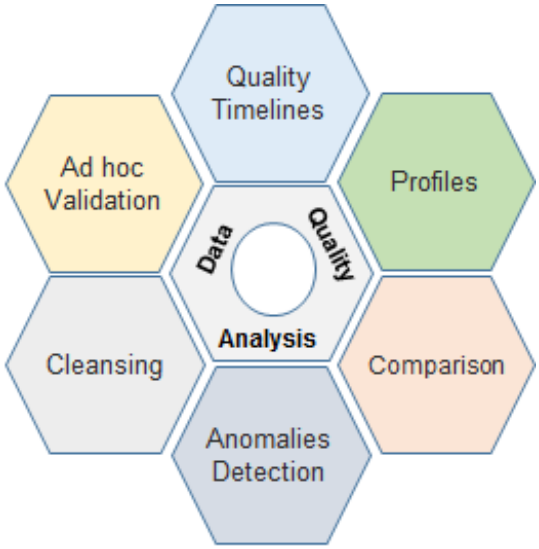

Figure 3

Data Validation and Quality Metering

Jumbune Data validation module provides deep insights into the quality of the data present on your Data Lakes. HDFS data validation is a generic data validation framework that

checks the data for anomalies and reports them with various details like type of violation, line number, file name containing maximum number of violation, most populated fields etc., it validates the data on the DFS based on custom defined set of rules as per the business. The rules can be in the form of null checks, data types and/or regular expressions expressing the business form of data. The data validation tasks can be scheduled and Data Quality Timelines are used to infer the health of the data over a period of time.

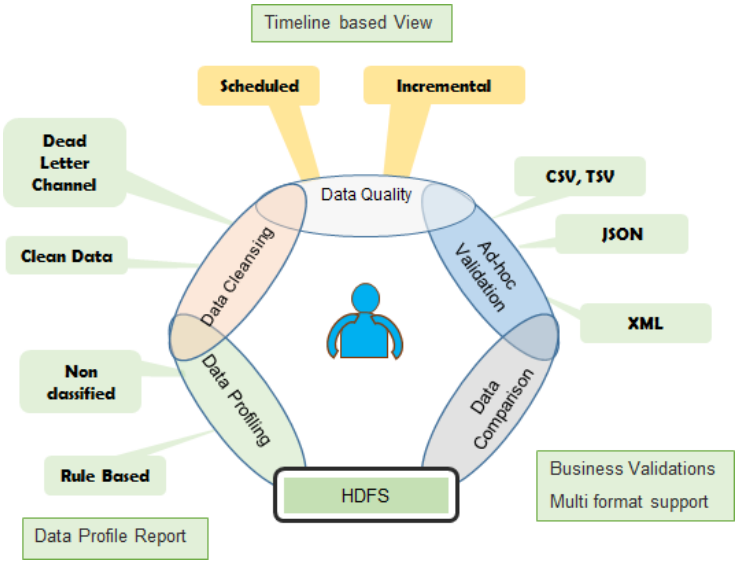

Figure 4

It supports various data formats such as plain text, CSV, TSV, JSON and XML documents as well.
Jumbune Data validation module provides following benefits

- *Real-time information of Data with changing dynamics.*
- *Faster predictive analysis based on trustworthy and accurate data*
- *Enhanced data quality enabling correct decisions and subsequent actions.*
- *Right insights from the minutest data.*
- *Reduce data testing and validation time.*

Data Cleansing

Another prominent offered feature in the space of Data Validation is data cleaning, also called data cleansing or scrubbing, deals with detecting and removing errors and inconsistencies from data in order to improve the quality of data. In analytics and Big Data testing the most tedious job is to cleanse the data.

Jumbune data cleansing feature not only cleanses the data in automated fashions while implementing business rule in form of null checks, data type check and regular expression but also provides Dead Letter Channel facility where inconsistent data is stored for future purpose. Jumbune also provides reports which depicts counts of clean and dirty filesets.

Data Profiling

Jumbune Data Profiling module helps in discovering business knowledge embedded in data itself. Jumbune uses structural approach for its deep analysis. Information obtained during data profiling such as data type, length, discrete values, uniqueness, occurrence of null values, typical string patterns can be used to discover problems such as illegal values, misspellings, missing values, varying value representation, and duplicates. HDFS data can also be profiled over a set of rules or without one to obtain some quick insights over the ingested data

Data Lake Pipelines – Integration of Jumbune

The below fig.5 portrays a typical scenario of data lake pipeline and how Jumbune data validation modules can be applied for data validation process.

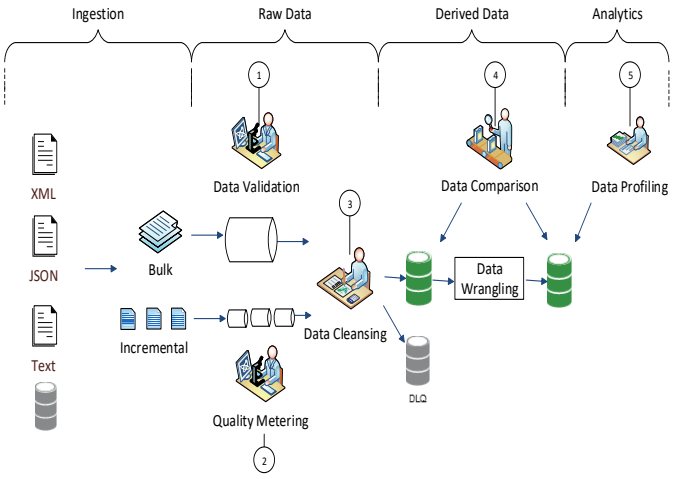

Figure 5

Jumbune is completely interoperable with most of existing enterprise Hadoop distribution running in Cloud, On Premise or Hybrid Infrastructure. All features offered by Jumbune works complementary with distribution and promises to deliver extended capabilities. A full functional copy of Jumbune can be download from its website. http://jumbune.com/