indium
Make Technology Work

Data Quality Validation

Making Sense of Big Data

White Paper

www.indiumsoftware.com

Big Data – The Emerging Trends

The market research firm Gartner defines Big Data as: "...high-volume, high-velocity and/or high-variety information assets that demand costeffective, innovative forms of information processing that enable enhanced insight, decision making, and process automation." Mordor Intelligence report predicts that global big data technology & service market will grow from $21.19 billion in 2017 to $77.58 billion by 2023, at an estimated CAGR of 24.15 per cent.

As a result of the adoption of Big Data technologies, conventional database systems are being replaced by horizontal database, columnar designs and cloud enabled schemas. But one area that remains evasive as of now is the quality of analysis due to the different mindset, skills and knowledge it needs.

Characteristics of Big Data

- Big Data, as the name suggests, refers to data that runs into terabytes or petabytes
- The four dimensions of volume, variety, velocity and veracity
- As data replacement happens at a rapid pace, processing needs to match the speed
- It can be unstructured and structured

Emerging Trends

McKinsey Global Institute's report, 'The Age of Analytics: Competing in a Data-Driven World' predicts that big data adoption will be further boosted by deep learning, AI and IoT technologies that are being adopted by organizations.

Some of the trends include:

- The growth of data stores such as Kudu and adoption of faster databases like MemSQL and Exasol have enabled speeding up Hadoop. Various SQL-on-Hadoop tools such as Hive, Impala, Phoenix, Presto and Drill, have enabled query accelerators to bridge the gap between traditional data warehouse systems and big data.
- It is not just Hadoop anymore. Big Data solutions are have a role to play in a variety of devices such as mobile apps, connected cars, wearables like FitBit, and Smart Meters. Therefore, Hadoop is being integrated with other big data technologies such as in-memory frameworks, data marts, discovery tools and data warehouses to deliver the data in a timely manner. Since data is culled from Hadoop and non- Hadoop sources, data and source agnostic technologies have become a must.
- Data lake strategy is gaining importance for centralized applications so that they can come together on a single central platform for better business decision making, innovation, improving profitability and gaining competitive edge.
- Data governance and security are under focus as hadoop becomes business-critical for organizations and components such as Apache Atlas, Apache Sentry and Apache Data are becoming key to data protection strategies.
- Cloud based hadoop deployments will be in demand as it will be cost effective even while enabling storing historical data at lower cost, thus ensuring accessibility and availability.
- As of now, Hadoop's primary attraction is in its scalability enabling the building of supersized data warehouses. But it is also a platform for executing machine learning algorithms that can help glean meaningful insights from vast and varied data for predictive analysis.
- According to big data experts, by the end of 2020, there will be 26 billion to 100 billion connected devices, and the focus of predictive analytics will turn on human interactions, transforming IoT to IoP.

Data Quality – Its Criticality

Data integrity is very critical for businesses. Garbage in/garbage out is the fundamental problem of poor data quality. This gets reflected in:

- Poor business decisions that can harm the business or not result in the expected growth
- Lack of credibility – poor data can raise serious concerns about credibility amongst customers, thereby driving them away and causing losses

The Need to Address Data Quality

Big Data insights help in decision making across business operations, including:

- Marketing
- Sales
- Customer support
- Service
- Fraud detection

Inaccurate and incomplete data can affect decisions related to:

- Targeting
- Segmentation
- Acquisition
- Cross-selling

This can result in increased cost of returned or undelivered materials, leading to lost opportunities and customer dissatisfaction. Unreliable data will also make managing risks, make strategic plans and complying with regulatory requirements challenging.

Data Quality Validation

Big Data is a new area of interest and yet to mature. Therefore, despite the presence of many tools, most organisations are yet to formalize a process to assure data quality management.

Typically, the validation happens with regard to tracking data load statuses, packet statuses, in an ad hoc or one-off manner.

It must establish DQV practices based on the following six principles.

Validate all Data

It is not enough to validate only sample data. There could be missing, duplicated records, or garbled records that would impact the accuracy of the reports.
To make complete validation possible, organisations must:

- Identify application areas that need to be validated
- Determine the level of granularity
- Set the time frame for data validation
- (days, months, quarters or years)
- Set the validation frequency (daily, weekly, monthly)

Primary Data Validation

Data should be validated at high level of granularity at this stage. Document or line item- level validation can be performed if a problem has been discovered and it hinders validation. A well-designed validation process can enable quickly refining the investigation, reducing the amount of data that needs to be analyzed at low granularity levels.

Prioritise Data Elements

Identify elements critical to the business that need to be analysed on priority.

Automate

Automated data validation ensures continuous execution and is integral to the design, quality assurance and production phases of the project lifecycle. This will also prove beneficial for BI implementation.

Know the Source

The source of data and how it is entered and stored in the source system database is important and requires understanding and commitment to data quality management, associated business processes, data entry modes, and data warehousing theory.

The key things to be checked include:

- The fields populated by the application
- The fields utilized in data warehouse reporting
- The frequency at which information is loaded into the data warehouse
- Time of data being archived out of the source system

Customise Validation

The frequency of validation, the parameters used for selection, and comparison object definition will vary depending on the business need. For this, validation should occur across the different layers of the data warehouse architecture being used and redundancy used to cross-validate data.

DQV - Benefits

Qualitative, accurate and healthy data ensure:

- Improved decision making
- Higher data accuracy
- Create a better business strategy with realistic goals
- Enhanced profit margins

Indium Software Approach

Given the importance of data in business decisions, Indium Software combines its testing expertise with understanding of Big Data to develop a process-oriented approach towards Data Quality Validation.

Test Process

The two aspects of the tests include:

- Integrity of the data type – the input data should remain the same throughout.
- Whether the multiple transformations of data is based on the business logic and if the logic is applied correctly.

Every row of the data has to be verified to validate it and therefore, the customer infrastructure is replicated using the right data stacks and tools.

When the data gets transformed, the testing includes:

- Whether data merging is as per the requirement.
- Check for the correctness of data types.
- Ensure that the row count is correct.
- Verify that the data categories are correct.
- Ensure that the numeric data is within the correct mean range.
- Check for missing values and rows.

The business logic is broken down into multiple steps and each step is checked to make sure the business logic application is correct and as expected.

Tools

To replicate the environment truthfully, a complementary tool is used. This avoids unnecessary errors and at the same time, helps detect any anomalies in the output. The variation, if any, is reported as a bug.

The testing is automated by scheduling it in a loop. Distributed testing helps to handle data overload.

Conclusion

Data volumes, velocities and varieties have created a deep need for data validation and verification. The right mind set, an understanding of the business reasons for the use of data and a meticulous testing approach are essential to ensure quality of data. The increasing challenge of storage, processing, and accessing the data, requires the right tools and technologies.

The right approach can ensure:

- Reduced cost and time.
- Integrated expertise and positioning of test environments to offer test data.
- Data security and amenability with data protection guidelines.
- Reduced cost and time.
- Integrated expertise and positioning of test environments to offer test data.
- Data security and amenability with data protection guidelines.

☎

INDIA

Chennai | Bengaluru | Mumbai
Toll-free: 1800-123-1191

USA

Cupertino | Princeton
Toll-free: +1 888 207 5969

UK

London

SINGAPORE

+65 9630 7959

Sales Inquiries

sales@indiumsoftware.com

General Inquiries

info@indiumsoftware.com