PES UNIVERSITY





Project Work Phase-II : UE18CS651
ESA

Project Title : Data Quality enhancement and monitoring framework for High Performance Computing Systems leveraging machine learning techniques.

Project Guide : Prof. Suresh Jamadagni

Student Name : Choragudi Praveen

SRN : PES1201802271

Problem Statement

- To recognize the **common** master data sets after gathering from incongruent systems in phase 1 of the project; sourcing the **clean** dataset and feed it to the **learning engine** and apply the generated model to auto-cleanse, suggest or run human-assisted/semi-automated scenarios.

- By doing this, we want to achieve considerable reduction in data quality eccentricities with on-the-fly fixing of about 70% of the master data entities, as recognized by Data Quality indicators and lively endorsements to fix about 20% of the persistent data sets.

User Profile

List of different users who are facing this problem:

- About 8 different persona (or role) deal with the system.
- Each of the persona have a different expectation of data quality, ranging all the way from a summarized level to very detailed, granular level.
- Senior management expects summarized data to be cross-mapped to the detailed data sets.
- Individual end users expect the data to be cross-mapped across data sets, and with 100% traceability .

Literature Survey

Year of Publish	Paper	Summary	Concluding Thoughts
2018	Big Data Platform for Enterprise project management digitization using Machine learning - an IEEE paper by Ruchi S and Dr Pravin Srinath	This paper benevolences the architectural design for efficient enterprise project management in current data scenario and analysis of the current enterprise project.	▪ Resistance of adoption of integrated platform for project management. ▪ Lack of integration of ML/DM techniques with existing project management framework.
2018	Data Integration Using Machine Learning - an IEEE paper by Marcus Birgersson, Gustav Hansson and Ulrik Franke	This paper presents a first version of a system that uses tools from artificial intelligence and machine learning to ease the integration of information systems, aiming to automate parts of it.	Fusing the results from the different models, and enabling continuous learning from an operational production system.
2019	Evaluation methods and decision theory for classification of streaming data with temporal dependence - a Springer paper by Albert Bifet, Jesse Read	This paper formalizes a learning and evaluation of predictive models on streaming data.	Not well-thought-out time-based dependence in data stream mining seriously enough when evaluating stream classifiers.
2019	A Study on Challenges and Opportunities in Master Data Management - a paper in IJDMS by Tapan Kumar Das and Manas Ranjan Mishra	This paper aims to provide a data definition of one master data for cross application consistency. The concepts related to Master data management in broader spectrum has been discussed.	▪ MDM is cross functional, it benefits from an organization that fosters collaboration between business and IT. ▪ A realistic research is needed to find out the actual state of MDM practices, current trend & their optimization.

PES UNIVERSITY




Proposed Solution

Machine Learning, Data Quality and Integration:

- ❑ Identifying the machine learning techniques that can be leveraged.
- ❑ Sourcing complete data.
- ❑ Apply Machine Learning techniques on entire data set, validate effectiveness.
- ❑ Create the API, integrate it with the data ingestion framework.

Solution Delivery:

- ❑ Streamlining the solution and deliver it to production.

PES UNIVERSITY




Why Your Solution is Better?

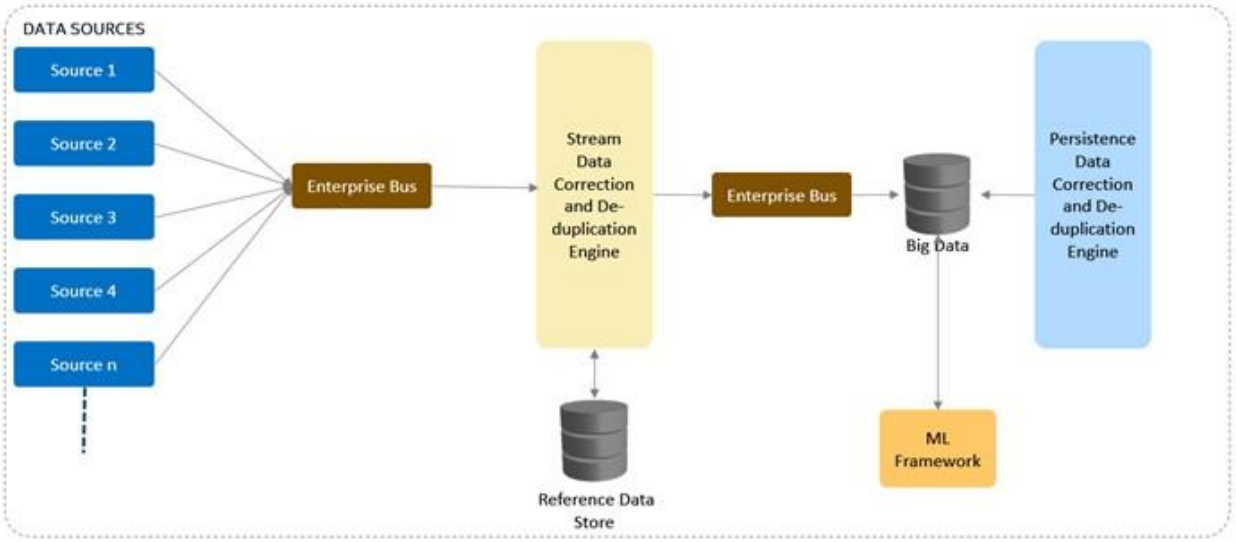
- No known occasions of such an application within the organization. A very multifaceted and a huge solution setting.
- One engine that corrects the data on the fly. This engine will be designed to fix less-data-intensive checks. The second engine fixes much more backsliding issues, that require data-intensive-operations.
- Both the system will leverage Machine learning, big data outline, data de-duplication and other data quality techniques to safeguard that data meets business expectations.
- Aggressive goals include:
  - Normalize and deduplicate static/master data using a wide-ranging variety of techniques including mapping, machine learning etc.
  - Drive incorporation ratio from 0% to about 80%, helping business with indicators that have not been conceivable beforehand.
  - Integrate the solution as a API with other modules such as ETL and monitoring framework, driving a seamless, end-to-end data lifecycle framework.

Technologies / Methodologies

- Python, NumPy - for data analysis
- Big Data technologies: Spark, Spark Streaming, HDFS, Hive, Pig, Scala (tentative), Beehive etc. - for data extraction and performing various operations on streaming and persistent data.
- Middleware: Apache Kafka - for real time data processing with distribution, performance and reliability, REST API - for communication with other modules.
- Spark MLlib - to make applied machine learning scalable and easy using it's built in tools such as ML Algorithms, Featurization, Pipelines, persistence and utilities.

The **Extreme Programming** software development approach is adopted as the proposed solution is research and exploratory in nature.

System Architecture



There are two mitigation points :

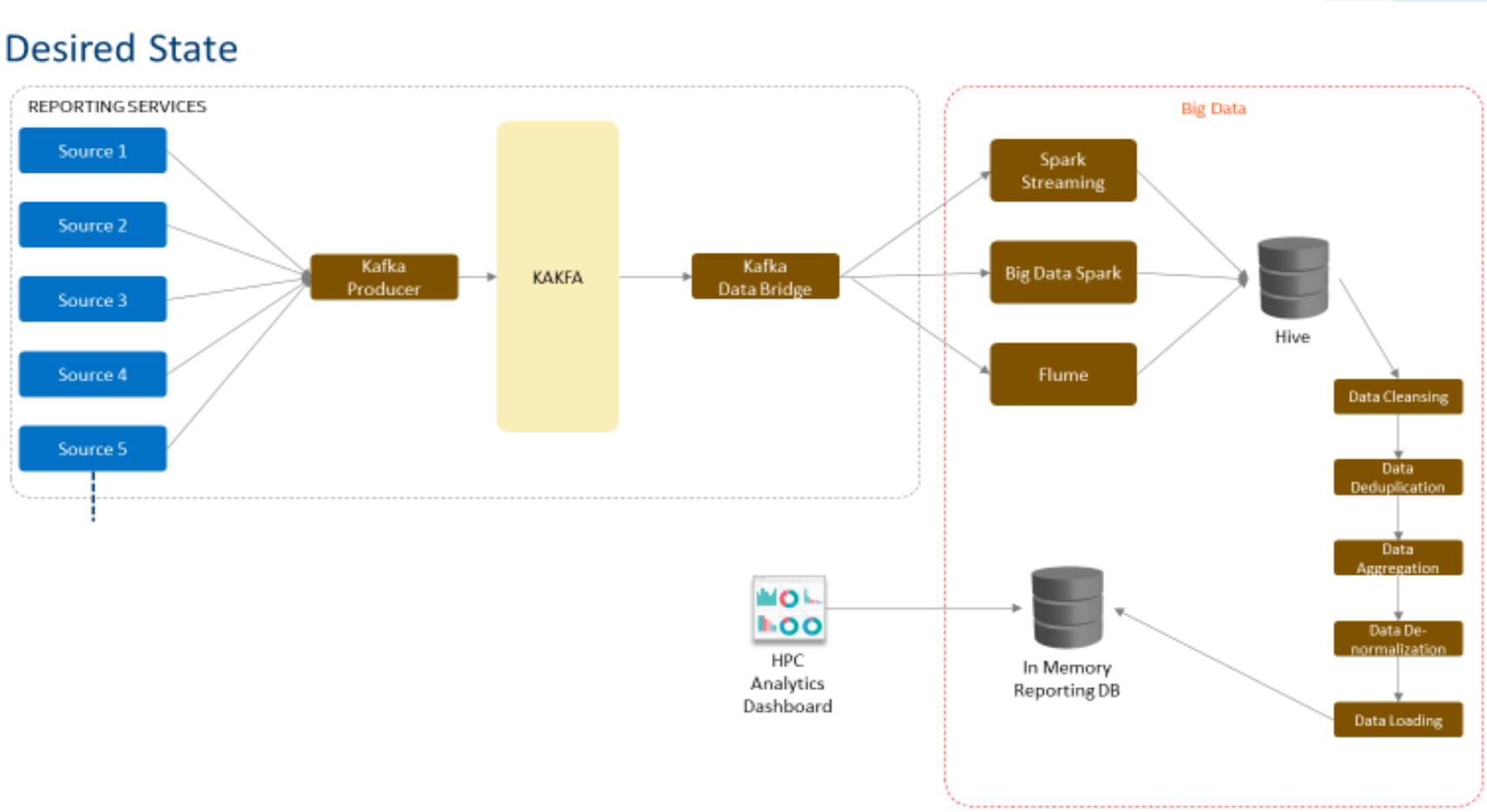
- One engine that corrects the information on the fly. This engine will be designed to fix less-data-intensive checks.
- The second engine fixes much more regressive issues, that require data-intensive-operations.

Both the system will leverage Machine learning, big data framework, data de-duplication and other data quality techniques to ensure that data meets business expectations.

PES UNIVERSITY




Use Case diagram with desired state



Test Strategy

Source	Model	Efficiency (with 30% of data)	Efficiency (with rest 70% of data)
ION	Linear Regression	84.88%	84.12%
AUTOOPS	Multiple Linear Regression	76.54%	76.13%
ICMS	Logistic Regression	94.63%	94.63%
FAAS	Linear Regression	89.68%	89.11%
MYHPC	Logistic Regression	90.24%	89.98%
COMPUTEBI	Linear Regression	95.56%	95.35%

- In sourcing data, we have followed a 30-70 approach, where we have sourced 30% of the data from a source and have trained the model, upon knowing the efficiency of the model leveraged we have extended the model to deal with rest 70% of the data.

Implementation Details

Source	ML model	Predictand	Predictor(s)
ION	Linear Regression	Maxrss	waittime
AUTOOPS	Multiple Linear Regression	Allocationcost	Vms assigned, cores
ICMS	Logistic Regression	demandDriverGroup-bgs’	Demand driver project id
FAAS	Linear Regression	Quarterly allocation cost	Current cost
MYHPC	Logistic Regression	Tier- normal/critical/soft	Utilization time
COMPUTEBI	Linear Regression	efficiency	Time on machine

Project RESULTS with Planned Effort Vs Actual Effort

KPI	Target	Current	Source	frequency
No of disparate sources	135	11	Legacy Datamart	Quarterly
No of clean datasets acquired	11	9	Big Data System	Monthly
ML technique identified datasets	11	6	Big Data System	Monthly
Overall efficiency of the proposed method on sourced data	100%	88%	Big Data System	Quarterly

PES UNIVERSITY

Thank You