PES UNIVERSITY
(Established under Karnataka Act No. 16 of 2013)
100-ft Ring Road, Bengaluru – 560 085, Karnataka, India

A Project Report On

Data Quality enhancement and monitoring framework for High Performance Computing Systems leveraging Machine Learning Techniques

Submitted in fulfillment of the requirements for the Project phase -2

Submitted by

Choragudi Praveen

PES1201802271

Under the guidance of

Prof. Suresh Jamadagni

Associate Professor

Jan- May 2020

FACULTY OF ENGINEERING AND TECHNOLOGY

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

PROGRAM M. TECH

FACULTY OF ENGINEERING

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

PROGRAM M. TECH

CERTIFICATE

This is to certify that the Dissertation entitled

Data Quality enhancement and monitoring framework for High Performance Computing Systems leveraging Machine Learning Techniques

is a bonafide work carried out by

Choragudi Praveen

PES1201802271

In partial fulfillment for the completion of 4th semester Project Work in the Program of Study MTECH in Computer Science and Engineering under rules and regulations of PES University, Bengaluru during the period Jan. 2020 – May. 2020. It is certified that all corrections/suggestions indicated for internal assessment have been incorporated in the report. The project report has been approved as it satisfies the 4th semester academic requirements in respect of project work.

Signature with date & Seal	*Signature with date & Seal*	*Signature with date & Seal*
Internal Guide	*Chairperson*	*Dean of Faculty*

Name and Signatures of the Examiners

1.

2.

3.

ACKNOWLEDGEMENT

I extend my deep sense of gratitude and sincere thanks to our chairperson Dr. M.R.Doreswamy (Founder, Chancellor PES University), Prof. Jawahar Doreswamy, ProChancellor of PES University and Dr. J. Suryaprasad, the Vice Chancellor of PES University for giving me an opportunity to be a student of this reputed institution.

I would like to thank PES University for providing me the educational background and conducive environment required for the completion of this project.

It is my Privilege to thank Chairperson of the Department Dr. Shylaja S S, Department of Computer Science and Engineering for her support and guidance for doing my Project.

I express my sincere gratitude to my guide Prof. Suresh Jamadagni for his valuable guidance and suggestion technically and logically for doing my Project work.

Finally, I would like to convey my regards to our faculty members, family and fellow mates for standing by throughout the period of my project completion.

Ch. Praveen

Choragudi Praveen
PES1201802271

Regd. Office:
Intel Technology India Pvt. Ltd.
Campus 4B, ECOSPACE
Outer Ring Road, Bellandur
Bengaluru 560103
India.

Tel: +91-80-2605 3000
FAX: +91-80-2605 5357
www.intel.in

Monday, July 29, 2019

To Whomsoever It May Concern

This is to certify that PRAVEEN CHORAGUDI (Employee Number: 11909576) is working in our company since Jun 19 2019 as a Graduate Intern Technical .

Thanking you,

for Intel Technology India (P) Ltd

Authorized Signatory

Table of Contents

1 Introduction 1
1.1 Problem Definition: 1
1.2 Generic Proposed Solution: 2
2 Literature survey and/or Studies Done 2
2.1 IEEE paper by Ruchi S and Dr Pravin Srinath 2
2.2 IEEE paper by Marcus Birgersson, Gustav Hansson and Ulrik Franke 3
2.3 a Springer paper by Albert Bifet, Jesse Read 4
2.4 An article published by Research Gate by Tapan Kumar Das 4
3 System Requirements Specification 5
3.1 High level block diagram of the Solution 5
3.2 Environment Used in the Project 6
3.2.1 Hardware Required 6
3.2.2 Software Required 6
3.2.3 Requirements for the Project 7
3.2.4 Constraints and Dependencies 8
3.2.5 Assumption 9
3.2.6 Use Case Diagrams for the requirements 9
4 Schedule 10
5 Design 11
5.1 Architectural Diagram 11
5.2 A sequence diagram or Interaction diagram 12
5.3 Data Flow Diagram 12
6 Implementation 13
6.1 Acquiring Clean Datasets 13
6.2 Identification of the ML technique to leverage on Clean datasets: 13
6.3 Sourcing Data: 14
6.4 API creation and Integration: 14
6.5 Solution Delivery: 14
7 Test Strategy: 14
8 Results & Discussions/Conclusions 15
9 Reference or Bibliography 16

1 Introduction

HPC Business is unable to obtain actionable indicators due to diverse data sources that are not integrated, and data is not of right quality.

- Information flows in from over 150+ projects into a central repository.
- The Key master data entities across the sources aren't standardized which leads to ineffective/non-actionable business indicators.
- Information from various sources needs to flow into a centralized DataMart in a summarized fashion. Information lacks traceability; hence data is less actionable.

Potential points of Failure:

- Unable to connect datasets to arrive a meaningful 360^0 perspective of data.
- Information drill down, data slicing and dicing is ineffective, and not 360^0.
- Problem finding/fact finding nearly impossible.
- The frequency of occurrence is many times in a day, for thousands of users dealing with billions of records per day.

Scope and its importance:

- Information sources integrated.
- Significant reduction in data quality deviations.

1.1 Problem Definition:

The data that is originating from the source systems is not standardized, integrated, deduplicated and cleansed leading to ineffective/non-actionable business indicators.

1.2 Generic Proposed Solution:

The data must be aggregated to identify the common master data sets; 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 significant reduction in data quality deviations with on-the-fly fixing of about 70% of the master data entities, as identified by Data Quality indicators and dynamic recommendations to fix about 20% of the persistent data sets.

2 Literature survey and/or Studies Done

2.1 IEEE paper by Ruchi S and Dr Pravin Srinath

Big Data Platform for Business project administration digitization using Machine learning

This paper presents the proposed architectural strategy and related conception enlightenment for resourceful initiative venture administration in current data setting. Also, an analysis of the current business project structures has been presented along with checks that need to be gabbed.

- They've proposed an EPM stage using Big Data and machine learning beliefs. The value of this design is the actual digitization of assignment related data and transactional records in various expositions available both within and outside the firm concerned.

- Standard project association tools are unconnected software fixing on few of project running parameters. Hence, they expressed six EPM limits for project debt using Big Data platform.

- This paper catalogues and addresses analysis violations of EPM in detail. This platform would improve the gap between the different supporters convoluted in headway of the probe.

Inadequacies in this paper:

- Absence of incorporation of ML/DM techniques with current scheme business outline.
- Opposition of implementation of combined platform for project management.
- Field dependence occurs for accessible procedures applied to edge projects making EPM multifaceted to scheme.

2.2 IEEE paper by Marcus Birgersson, Gustav Hansson and Ulrik Franke

Data Incorporation By means of Machine Learning

This paper boons a main diversity of a system that habits tools from artificial intelligence and machine learning to affluence the amalgamation of information structures, steering to mechanize fragments of the situation.

- The Expanse model, in maximum cases has the utmost reassurance altitudes for mappings and the uppermost notches as the model understands paths that have been mapped previously, and maps them over.

- The Data assessment model in most cases achieves shoddier as the model was time-honored to extricate amid data tenets of unlike type, preferably trajectories or libretti.

- The models in general can associate tracks in the schedules, they can be rummage-sale to save period in the amalgamation process.

Snags in this paper:

Blending the results from miscellaneous models, and allowing incessant learning from a operational structure.

2.3 a Springer paper by Albert Bifet, Jesse Read

Appraisal approaches and choice philosophy for sorting of streaming data with chronological need

This paper supposedly examined assessment of classifiers on streaming data with sequential reliance. It suggests that the habitually acknowledged facts stream grouping whereabouts, such as taxonomy truth and Kappa figure, slump to detect situations of unfortunate recital when temporal need is existing, so they should be unused as only performance pointers.

- Classification correctness can be deceptive if used as a replacement for assessing alteration finders with datasets having temporal necessity.
- The decision theory for flowing data classification with temporal dependence and change new practice for data stream cataloguing that takes temporal dependence.
- Proposed a joint degree for cataloging enactment, that considers temporal dependence, and use it as the performance ration in classifying flowing data.

Concluding Thought:

Not well-thought-out time-based reliance in stream data mining when evaluating stream classifiers.

2.4 An article published by Research Gate by Tapan Kumar Das

Confronts and Prospects in Master Data Supervision

This paper provides data definition of one master data for cranky application steadiness. Master data management in wider range has been conferred.

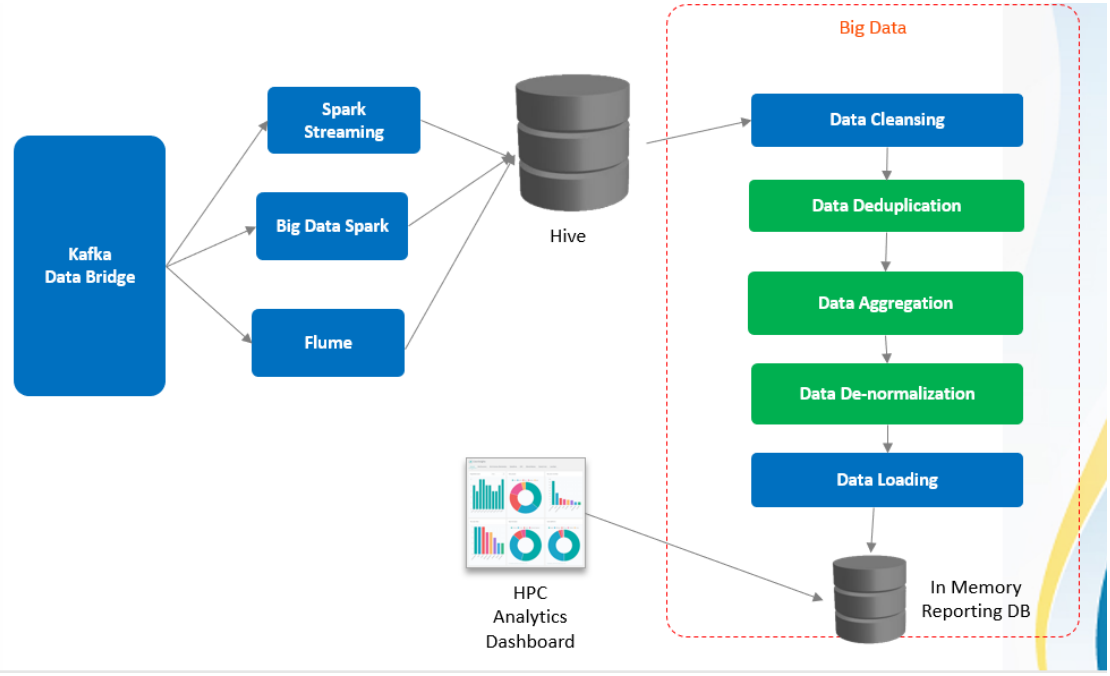
- The Master data requires actualities labelling of a commerce item, with more than one aspect. Entity definitions based on principal data gives steadiness in business and data honorableness when various IT systems across an association classify same data field differently.
- Business analytics becomes pertinent and modest information from big data depositories, but unpredictable data origins prevail. Data quality distress effective know-how and yield imprecise hearsays.

Concluding Thoughts:

- MDM is cantankerous working; it paybacks from a business that cultivates treaty amongst business besides IT.
- Research is needed to find out the authenticity of MDM tries.

3 System Requirements Specification

3.1 High level block diagram of the Solution



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.

3.2 Environment Used in the Project

3.2.1 Hardware Required

The details of the VM allocated in order to run our solution are:

Operating System	Windows Server 2016 Data Center
Processor	Intel® Xeon® CPU E5 – 2698 v4 @ 2.20GHz (4 processors)
Installed RAM	16.0 GB
System type	64-bit operating system, x64-based processor
Storage	79.9GB

3.2.2 Software Required

- 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.

3.2.3 Requirements for the Project

Functional Requirements:

F1. Given the input source location, the elastic dump should get the index file based on the current date and time of the execution.

F2. Once the index file gets downloaded from the source, all the job history files shall be dumped upon proper parsing.

F3. Before dumping of the files, we need to make sure that the Apache Kafka service is running in the specified server location which is given as a parameter to the producer.

F4. The flow of data from the disparate sources isn't limited by time. Henceforth, a time limit can't be defined for the incoming data. To achieve this, the producer must be continually running in the periodic intervals to avoid data loss.

F5. Consumer also should be running parallelly so that no data loss happens in between. The data types of the fields in the data are to be adjusted in order to make more sense of storage.

User Interface:

UI isn't planned as the part of the solution and we may have an UI after the proposed solution is in the production successfully as an additional feature.

Non-Functional Requirements:

NF1. We are creating staging tables because it helps us to identify any data type mismatches.

NF2. The data processing is done in batches so that we need to store for further phases of analysis and to ease the computing as data is voluminous.

NF3. The data which is used as training data needed to be stored separately and should feature to add more data in the process of training the model.

3.2.4 Constraints and Dependencies

Constraints:

- As the data is flowing in from disparate systems, it should be parsed and stored in the database. The risk involved here is to be able to parse voluminous data avoiding memory related errors.
- For the real time data processing, data loss is expected. So, it must be taken care by proper cluster management or by deploying adequate consumer groups or by creating multiple topics in apache Kafka.
- Getting the data from the database consuming large amount of time irrespective of parallel pipeline implementation.
- Unable to analyze the expected amount of data volume because of the hardware inconsistencies hence, periodical analysis is implemented.

Dependencies:

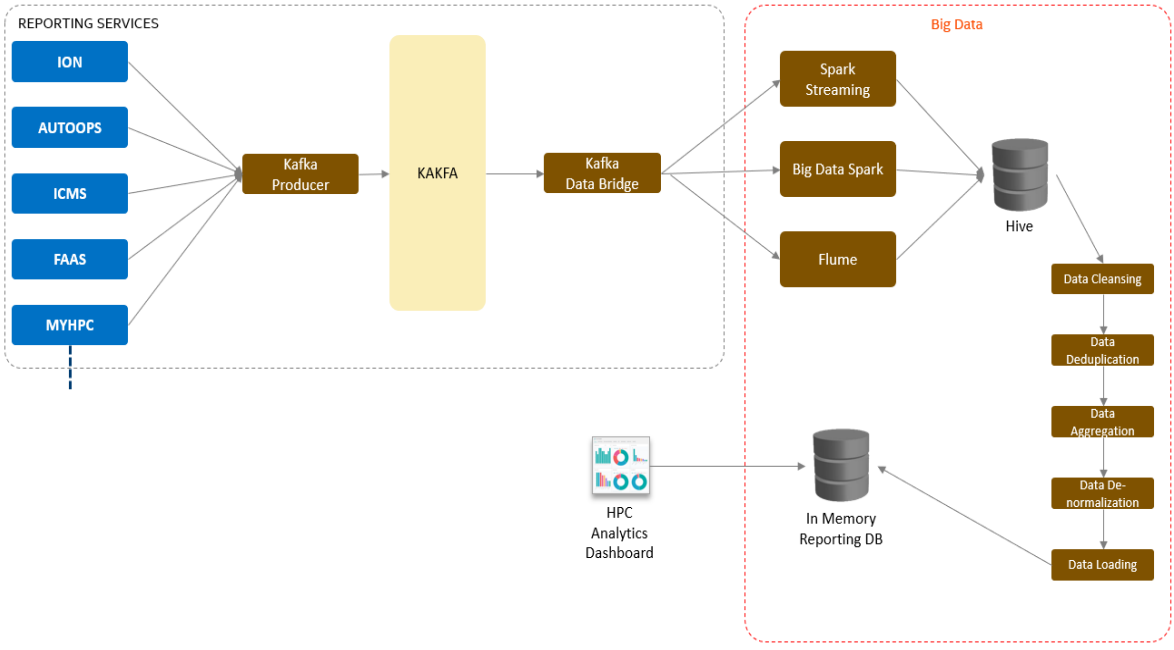
In order to proper functioning of the system, it is essential that the data sources shall not be down which leads to inactivity or a halt in the data ingestion pipeline.

3.2.5 Assumption

- The addition of the more features within a data source shall not be too frequent as it affects the enactment of the learning engine.
- To build and use models with high traceability prediction with good accuracy.

3.2.6 Use Case Diagrams for the requirements

- 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 with 100% traceability.



- Aggregate and integrate data from extremely diverse structured and unstructured data sources.
- Standardize and deduplicate static/master data using a wide variety of techniques including mapping, machine learning etc.
- Drive integration ratio from 0% to about 80%, helping business with indicators that have not been possible before.
- Integrate the solution as an API with other modules such as ETL and monitoring framework, driving a seamless, end-to-end data lifecycle framework.

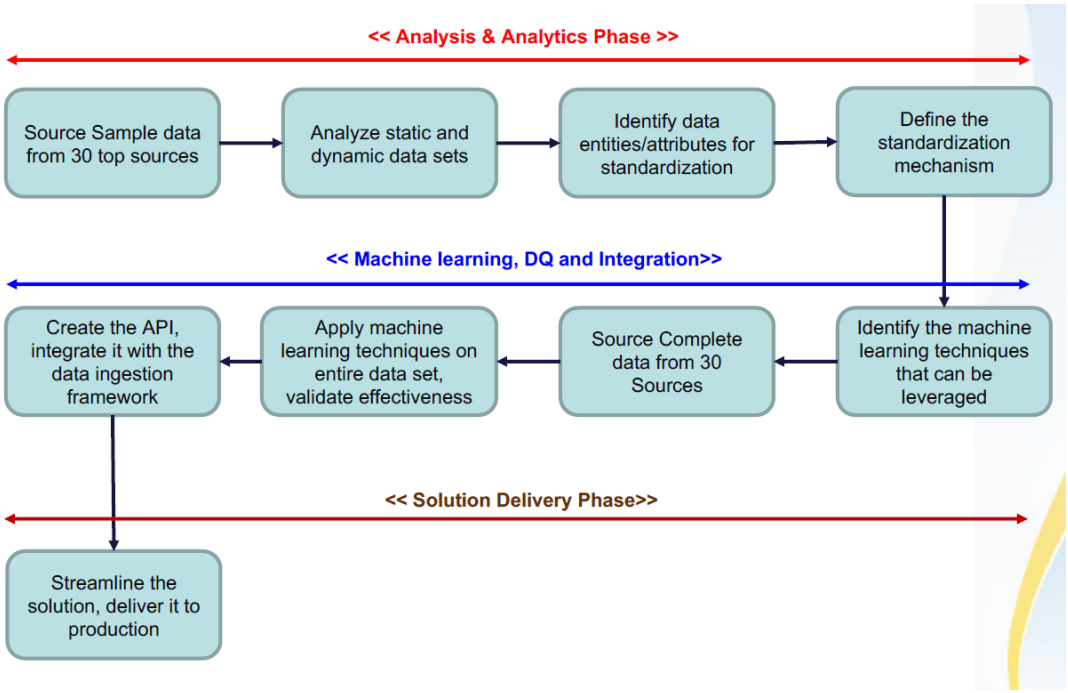
4 Schedule

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

Table: Results showing planned vs actual effort.

5 Design

5.1 Architectural Diagram



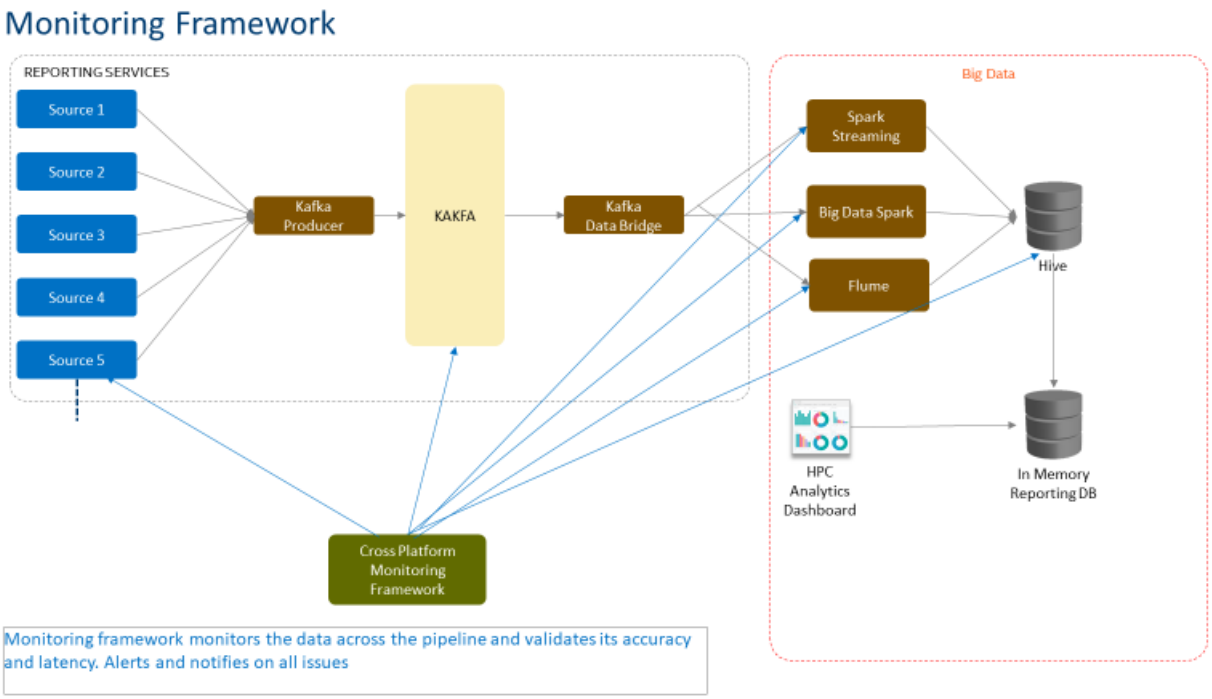
Analysis & Analytics Phase:

- Sourcing **sample** data from top source.
- Analyze static and dynamic data sets.
- Identify data entities/ attributes for standardization.
- Defining the standardization mechanism.

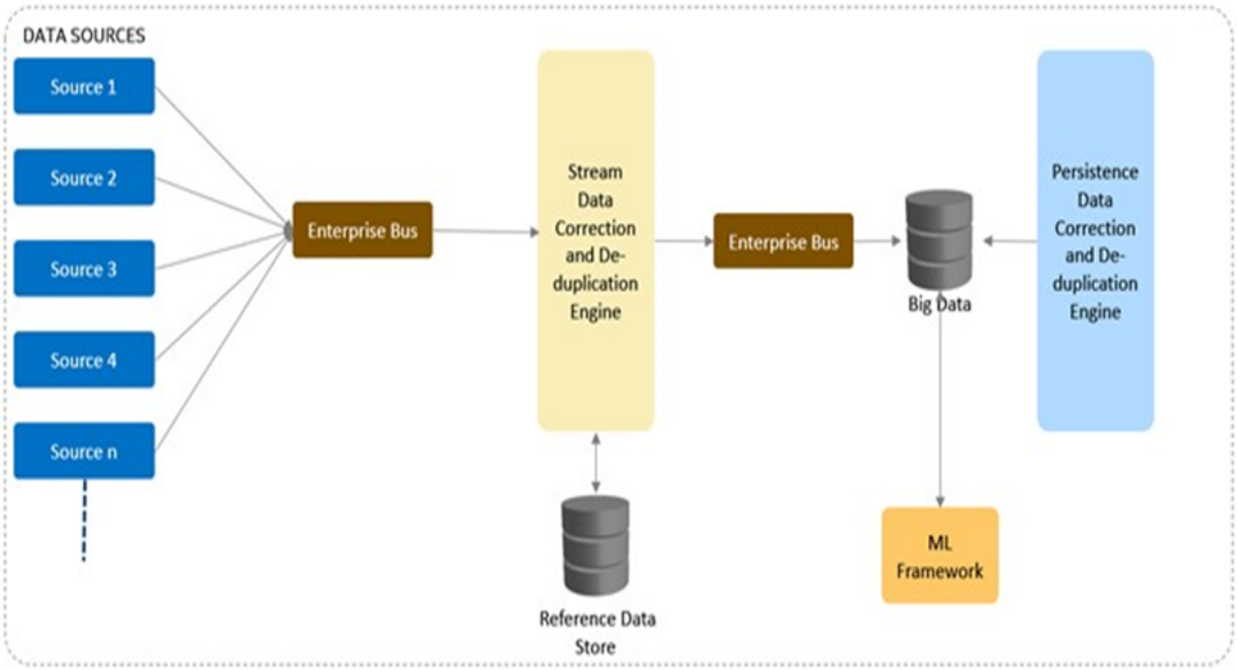
Machine Learning, Data Quality and Integration:

- Acquiring Clean Data Sets
- Identification of the ML technique to leverage on Clean datasets
- Sourcing of complete data.
- Creating the API and integrating it with the data ingestion framework.

5.2 A sequence diagram or Interaction diagram



5.3 Data Flow Diagram



The project started with an assumption of 135 sources, but at this point of time we are completely sourcing data from 11 sources. They are:

- ION
- NBCALENDAR
- NBFLOW
- AUTOOPS
- ICMS
- ECDEBUG
- FAAS
- CLOUDP
- MYHPC
- COMPUTEBI
- NBCONSOLE

6 Implementation

6.1 Acquiring Clean Datasets

Data preprocessing phase which started in Phase 1 of the project has achieved clean dataset for 6 sources and remaining 5 sources had clean data sets in phase 2 of the project.

6.2 Identification of the ML technique to leverage on Clean datasets:

The clean datasets that are obtained from phase 1 of the project were identified so as what ML techniques to be leveraged. The details are:

- ION – Linear Regression
- AUTOOPS – Multiple Linear Regression
- ICMS – Logistic Regression
- FAAS – Linear Regression
- MYHPC – Logistic Regression
- COMPUTEBI – Linear Regression

6.3 Sourcing Data:

- From the disparate sources, we have identified key master data entities from each source.
- By completing end-to-end data ingestion pipeline leveraging big data technologies like spark, apache-Kafka for stream data.
- Achieved clean datasets for 6 disparate sources.

6.4 API creation and Integration:

Creating restful API in order to integrate with the data ingestion framework is WIP (work in progress).

6.5 Solution Delivery:

- The goal that set is to CREATE A MASTER DATASET. This can only be achieved with implementing the proposed framework to all the 135 sources.
- But, for the fulfilment of the MTech project, I am planning to try out creating a mini Master dataset from the 6 disparate sources so far.

7 Test Strategy:

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.

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%

8 Results & Discussions/Conclusions

Source	ML model	Predictand	Predictor(s)
ION	Linear Regression	Maxrss	waittime
AUTOOPS	Multiple Linear Regression	Allocationcost	Vms assigned, cores
ICMS	Logistic Regression	demandDriver Group-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

9 Reference or Bibliography

[1] “http://activemq.apache.org/”

[2] “http://avro.apache.org/”

[3] “Cloudera’s Flume, https://github.com/cloudera/flume”

[4]http://developer.yahoo.com/blogs/hadoop/posts/2010/06/enabling_hadoop_batch_processi_1/

[5] Efficient data transfer through zero copy: https://www.ibm.com/developerworks/linux/library/jzerocopy/

[6] Facebook’s Scribe, http://www.facebook.com/note.php?note_id=32008268919

[7] IBM Websphere MQ: http://www-01.ibm.com/software/integration/wmq/

[8] http://hadoop.apache.org/

[9] http://hadoop.apache.org/hdfs/

[10] http://hadoop.apache.org/zookeeper/

[11] http://www.slideshare.net/cloudera/hw09-hadoop-baseddata-mining-platform-for-the-telecom-industry