A Framework to Maintain Data Relevance in a Data Lake

Tejas Dharamsi, Satyanarayana Srinivas

Sadiur Rahman, Rupam Rai

PES UNIVERSITY

Centre for Cloud Computing and Big Data

Content

- Overview
- Research Work
- Motivation
- Implementation
- Results
- Future Work
- Acknowledgment
- Biographies

Content

- Overview
- Research Work
- Motivation
- Implementation
- Results
- Future Work
- Acknowledgment
- Biographies

Overview

- Data Lake, accepts data from any source, in any format and arriving at any time.
- Data Lake will eventually have to deal with issues like data quality, validity of data, and its misuse.
- In our approach, various types of data were generated to meet the entry criteria of the Data Lake.

Overview

- The associated header was parsed to recognize the data source, validity of subscription, category of data, type & size of data, and duration for which the data should be persisted.
- Primary Node recognizes and directs the input data stream based on secondary attribute and name for data classification and storage.
- The framework avoids indiscriminate dumping of data in to the Data Lake by using subscription, hierarchy and period of validity of the data

Overview

- Data has been segregated and stored based on internal logical separation, which eliminate the eventuality of Data Swamp and help maintain the relevance of data.

Content

- Abstract
- Research Work
- Motivation
- Implementation
- Results
- Future Work
- Acknowledgment
- Biographies

Research Work

- Data, as we know, is a combination of qualitative and quantitative variables.
- The indiscriminate use of data with no quality control around it to keep the data relevant ends up in a Data Swamp.
- The maintenance of data quality is an important task,
  - Down stream analysis
  - Effective decision making

Research Work

- With Big Data taking stage,
  - Logical movement from databases to Big Data.
  - Will have to go through the collection process, in real time for analysis
- Certain aspects of data will continue to get the attention of the scientific community like –
  - Data quality, timeliness, reliability, currency, completeness and relevance of the data

Research Work

- As and when Big Data is drawing the attention, challenges are being looked at closely by engineers and scientists
  - The nature of Big Data is placing less and less relevance on what happened yesterday and more emphasis is being placed on what is happening now.

Research Work

- Data Lake being the concept for storing Big Data in a repository in:
  - Its native form
  - Will be worked upon only when needed
  - A storage place for all dates in different shapes and forms

Research Work

- In our paper, we have:
  - Designed a framework to accept all types of data in its native form.
  - Deal with the privacy and security issues in the Data Lake.
  - A method to preserve data relevance.
  - Subscription phenomenon for the services in Data Lake to avoid degradation in data quality
  - Use of commodity hardware

Content

- Abstract
- Research Work
- Motivation
- Implementation
- Results
- Future Work
- Acknowledgment
- Biographies

Motivation

- Big Data has become a reality, with many platforms of analytics being done for various problems
- Hadoop, HDFS and MapReduce is being used for number crunching and the benefits are being realized.
- Quality and relevance of data is still a challenge

Motivation

- Use cases are being formed and tried, for various scenarios
- Large volumes of data are being analyzed with out compromising on the critical data
- Data Lake is replacing Data Warehouse and we have technologies to do this
- Future analysis of data will be dependent on data quality and relevance

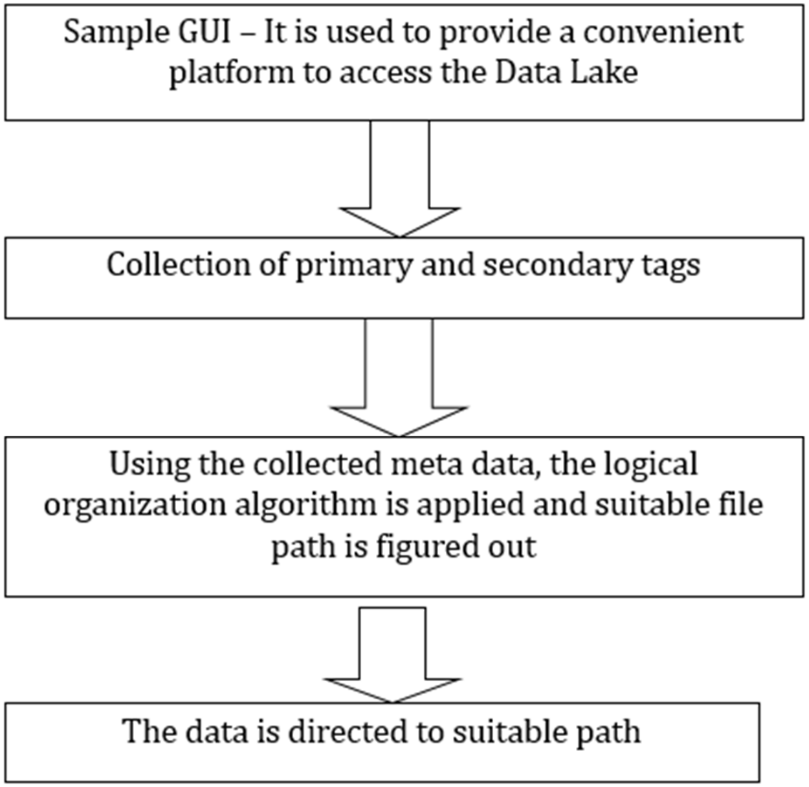
Content

- Abstract
- Research Work
- Motivation
- Implementation
- Results
- Future Work
- Acknowledgment
- Biographies

PES UNIVERSITY

Centre for Cloud Computing and Big Data

Implementation – Fig. 1.



Sample GUI – It is used to provide a convenient platform to access the Data Lake

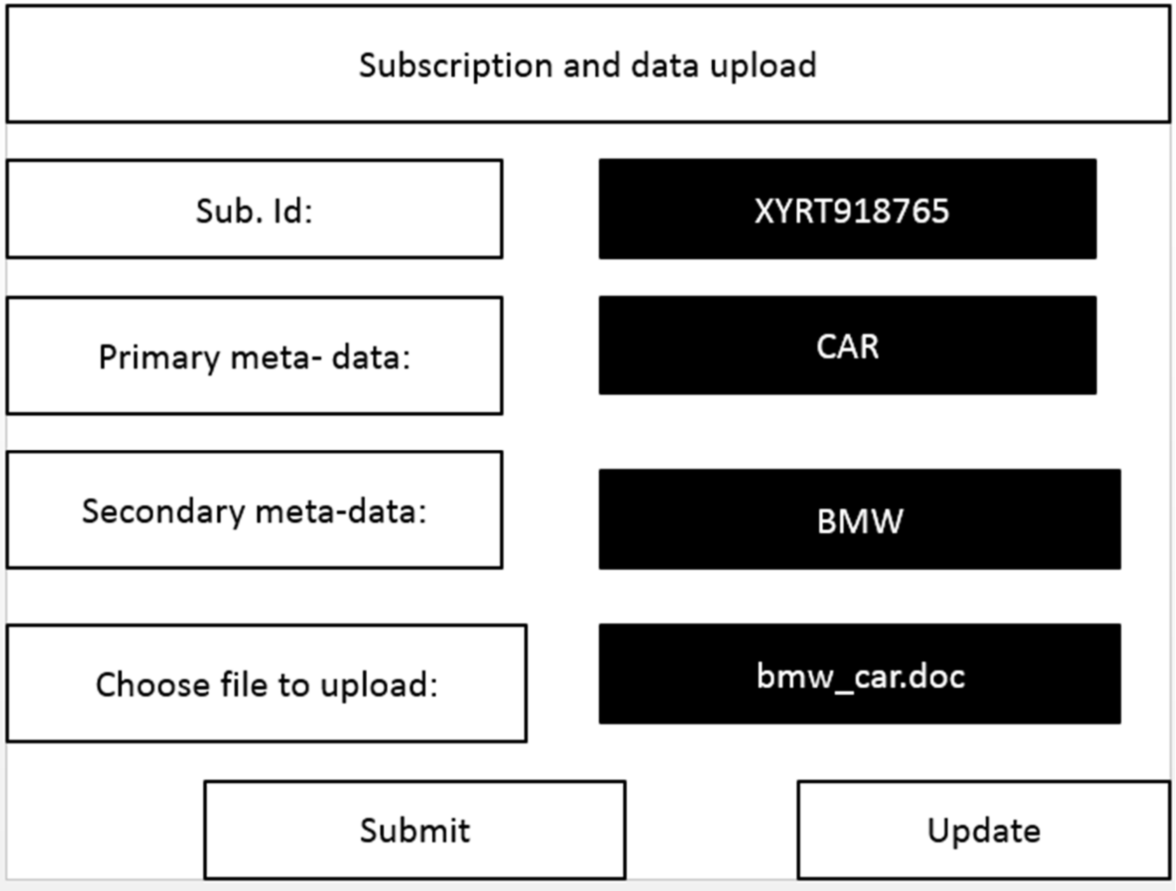
Collection of primary and secondary tags

Using the collected meta data, the logical organization algorithm is applied and suitable file path is figured out

The data is directed to suitable path

Fig. 1. Flow of data in to the Data Lake

Implementation – Fig. 2.



Subscription and data upload

Sub. Id:	XYRT918765
Primary meta- data:	CAR
Secondary meta-data:	BMW
Choose file to upload:	bmw_car.doc

Fig. 2. Shows the use of primary and secondary meta-data

Implementation – Fig. 3.

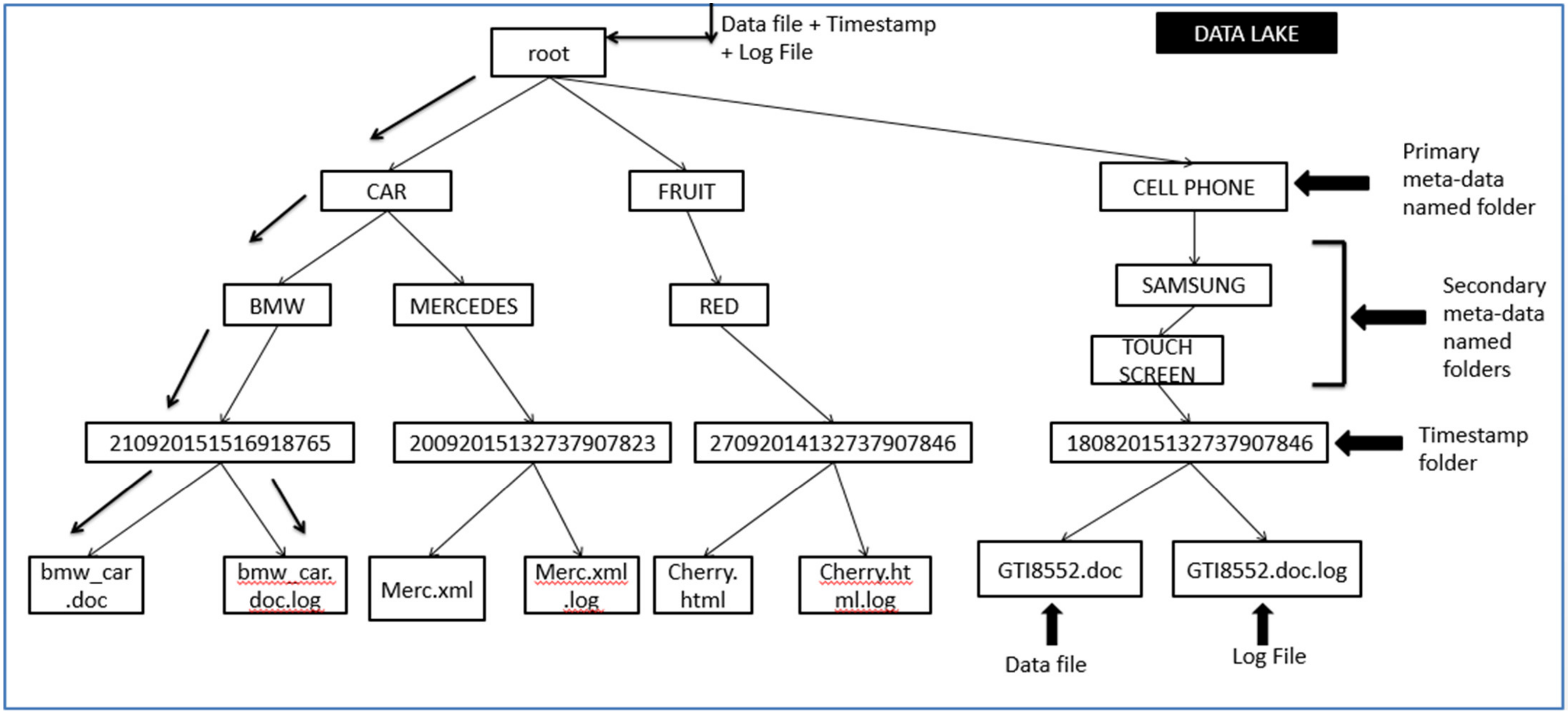


Fig. 3. Shows the organization and clustering of data

PES UNIVERSITY

Centre for Cloud Computing and Big Data

Content

- Abstract
- Research Work
- Motivation
- Implementation
- Results
- Future Work
- Acknowledgment
- Biographies

PES UNIVERSITY

Centre for Cloud Computing and Big Data

Results

- Big Data has been treated in real time
- Some aspects of making data has been cleared achieved and they are:
  - Identification of the of data,
  - Can arrive at any time,
  - Data type and form,
  - Identifying and recognizing where to be stored with the help of primary and secondary meta-data,

Results

- Use of commodity hardware to store the incoming data
- Subscription to eliminate indiscriminate dumping of data,
- Avoiding privacy and security challenges by identifying subscriber
- Since we have multiple entry points in to the Data Lake, the model has been delivery to scale up

Content

- Abstract
- Research Work
- Motivation
- Implementation
- Results
- Future Work
- Acknowledgment
- Biographies

Future Work

- Archival process of out dated data
- Implementing the variety of retrieval processes and a measure their efficiencies.
- Scaling the storage of data in to various file systems
- Compare and contracts against Hadoop based Data Lake with this plutonic implementation of Data Lake
- Implementation of Data Lake clusters to mimic the data marts
- Use HDFS and MapReduce on content of the Data Lake

Content

- Abstract
- Research Work
- Motivation
- Implementation
- Results
- Future Work
- Acknowledgment
- Biographies

Acknowledgment

- The authors would like to thank PES University, Bangalore, India for their continued support in carrying out this research and in publishing this paper at the conference. The authors thank Prof. D Jawahar, Prof Nitin V Pujari, Dr. K N B Murthy and Dr. K S Sridhar, for their support and encouragement. Finally authors thank Cloud Computing and Big Data Center at PES University, where the work was carried out

PES UNIVERSITY

Centre for Cloud Computing
and Big Data

Content

- Abstract
- Research Work
- Motivation
- Implementation
- Results
- Future Work
- Acknowledgment
- Biographies

Biographies

- **Tejas Dharamsi** is currently pursuing his Bachelors Degree in Computer Science Engineering at PES Institute of Technology, Bangalore India. He is currently Member Technical Staff at Ordell Ugo an environment to foster undergraduate research. He has been a student developer for Freenet Project Inc. at Google Summer of Code 2014 and is also the Google Student Ambassador, India - 2014 for PES Institute of Technology. He was a research intern with Carnegie Mellon University Electrical and Computer Engineering Department during the summer of 2015.

Biographies

- **Satyanarayana Srinivas**, a Research Scholar in PES University, his area of research interest is in Data Mining & Analytics, and Big Data Analytics with Visualisation, teaches Undergraduate and Post-Graduate students and guides them in research. His keen interests lies in fact that data explosion has happened and needs to be contained to study the hidden nuggets. He has been in the industry for 25 years and has led a number of multi-million dollars deals with Fortune 500 companies.

Biographies

- **Sadiur Rahman** is currently pursuing his Bachelors Degree in Computer Science Engineering at PES Institute of Technology, Bangalore India. He is interested in Data Mining, Big Data and its analytic. He is currently a Mozilla Firefox Student Ambassador for P.E.S Institute of Technology. He is also the main web developer for Mera Medicare, a U.S based pharmaceutical firm

Biographies

- **Rupam Rai** is currently pursuing her Bachelors Degree in Computer Science Engineering at PES Institute of Technology, Bangalore India. She is interested in Data Mining, Big Data and its analytic. Previously, an intern at S.A.I.L C&IT department.