WIPRO
Applying Thought

Real-Time Data Access Using Restful Framework for Multi-Platform Data Warehouse Environment

Pon Prabakaran Shanmugam,

Principal Consultant,
Wipro Analytics practice

Table of Contents

Abstract

The IT departments in enterprises see a lot of value in building service-oriented architecture around their data warehouse environment to empower their internal customers. The arrival of the Internet of Things (IoT) introduced a new deluge of data getting processed and used for analytics. With more data getting processed and stored, the need for multi-platform data warehouse environment has emerged. The volume, velocity and variety of data and its potential use for the organic growth of the business elicited the data platforms growing bigger.

Today, data warehouse environment in organizations are at the threshold of fulfilling diverse use cases and provide data to broad users across the spectrum like business applications, business intelligence, data analysts, data scientists, etc.

Real-time data ingestion and extraction need to be easier with or without the involvement of IT. With the availability of features like text analysis, pattern matching in analytical platforms, REST as a framework is a great vehicle to carry and retrieve data from the data process and storage engines.

This paper addresses how RESTFUL framework becomes a cost effective solution to achieve the mounting need to serve data in real-time.

Introduction

The heavy dependence on Extract, Transform and Load (ETL), and business intelligence tools has created some fatigue among business users. It takes multiple iterations and a long wait for businesses to get the

data that they need. The emergence of simple but efficient open source frameworks like REST enable fast movement of data using most popular web protocols.

Internet of Things (IoT) and Big Data

Internet-enabled computer embedded chips in products and devices are used primarily for data-gathering, offering enterprise-level details on everything from how efficiently their machines are running to the purchasing habits of their consumers.

Without the proper data-gathering in place, it will be impossible for businesses to sort through all the information flowing in from these embedded sensors. What that means is that, without analytics on the Big Data being captured, the Internet of Things can offer an enterprise only little more than noise[i].

Emergence of Multi-Platform Data Warehouse Environment

The 21st century was the period which marked the emergence of data warehouse as a science. The need to process and store data got traction with the business finding its usage. With more and more data getting processed, data appliances became popular. With the arrival of Internet of Things, data collection and processing got a new definition as the amount

of data being collected increased exponentially.

The need to build multiple platforms to process and store data has hit the organizations. With the introduction of architectural principles like Teradata Unified Data Architecture (UDA), there are a lot of options to build a true multi-platform data warehouse environment. It is possible to store data of any size. Data Lake gives the options of storing data as it comes and in any data format. A combination of platform which is interconnected gives the facility to move data between platforms. There is now an option to perform insights on data in real-time. Tools like Teradata Query Grid, helps to move data between platforms and also has features to retrieve data from different platforms without the user knowing where the data was stored.

The volume and variety of data is directly correlated to the number of components to process. The days of conventional batch processing and canned analytics don't satisfy the new type of users who use this data. That's why organizations are looking out for non-formal ways to integrate, store and access data. The Open Source RESTFUL framework is one of the technologies which facilitate the ease of data integration and extraction[ii].

RESTFUL Web services:

REST defines a set of architectural principles by which one can design web services that focus on a system's resources[iii]. Major advantages being:

• Use HTTP methods explicitly

• Be stateless

• Expose directory structure-like URIs

• Transfer XML, JavaScript Object Notation (JSON) or both

Building Real-Time Data Access with RESTFUL Framework

Given that multi-platform data warehouse environments have different work load capabilities, looking for real-time data ingestion and extraction becomes more difficult. Assume that there is a requirement to load and access unstructured data in real-time into multi-platform data warehouse environment. Being unstructured data, it makes sense to load into Hadoop (which is basically meant/good for batch processing). After the data is cleansed and ready for integration, it is meaningful to load the cleansed data into EDW or IDW to make it more efficient for the need for real-time access.

The web HDFS (Hadoop Data File Store) feature which is being offered by Apache or Hortonworks could be used for real-time data ingestion into HDFS. Also, the REST API can easily communicate to the Hadoop clusters. The file read and file write calls are redirected to the corresponding data nodes. It uses full bandwidth of the Hadoop cluster for streaming data[iv].

BIG DATA
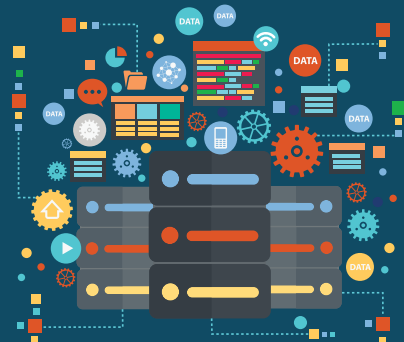
Conceptual Representation of Using REST for Multi-Platform Data Warehouse Environment
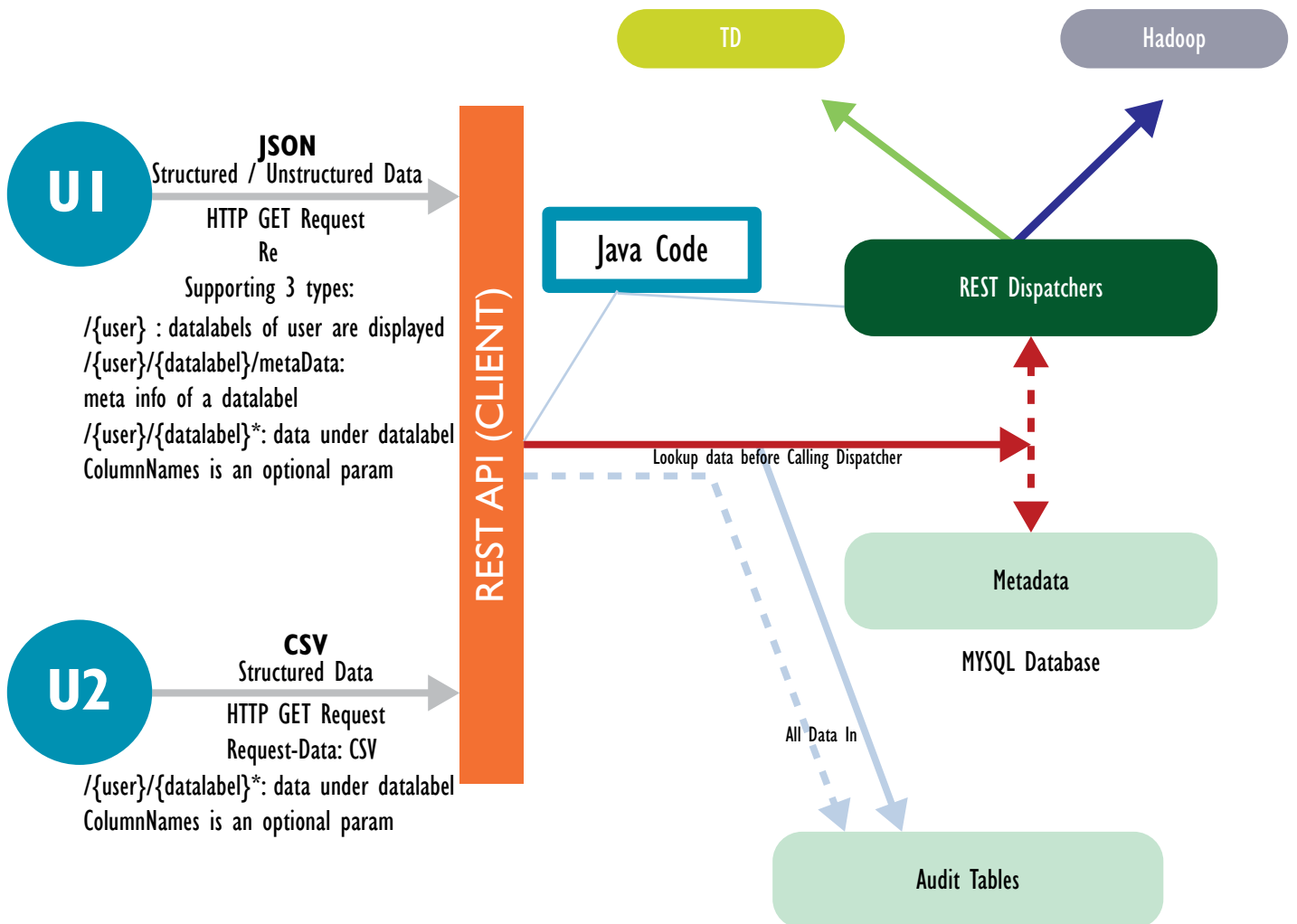
End Users – Data In
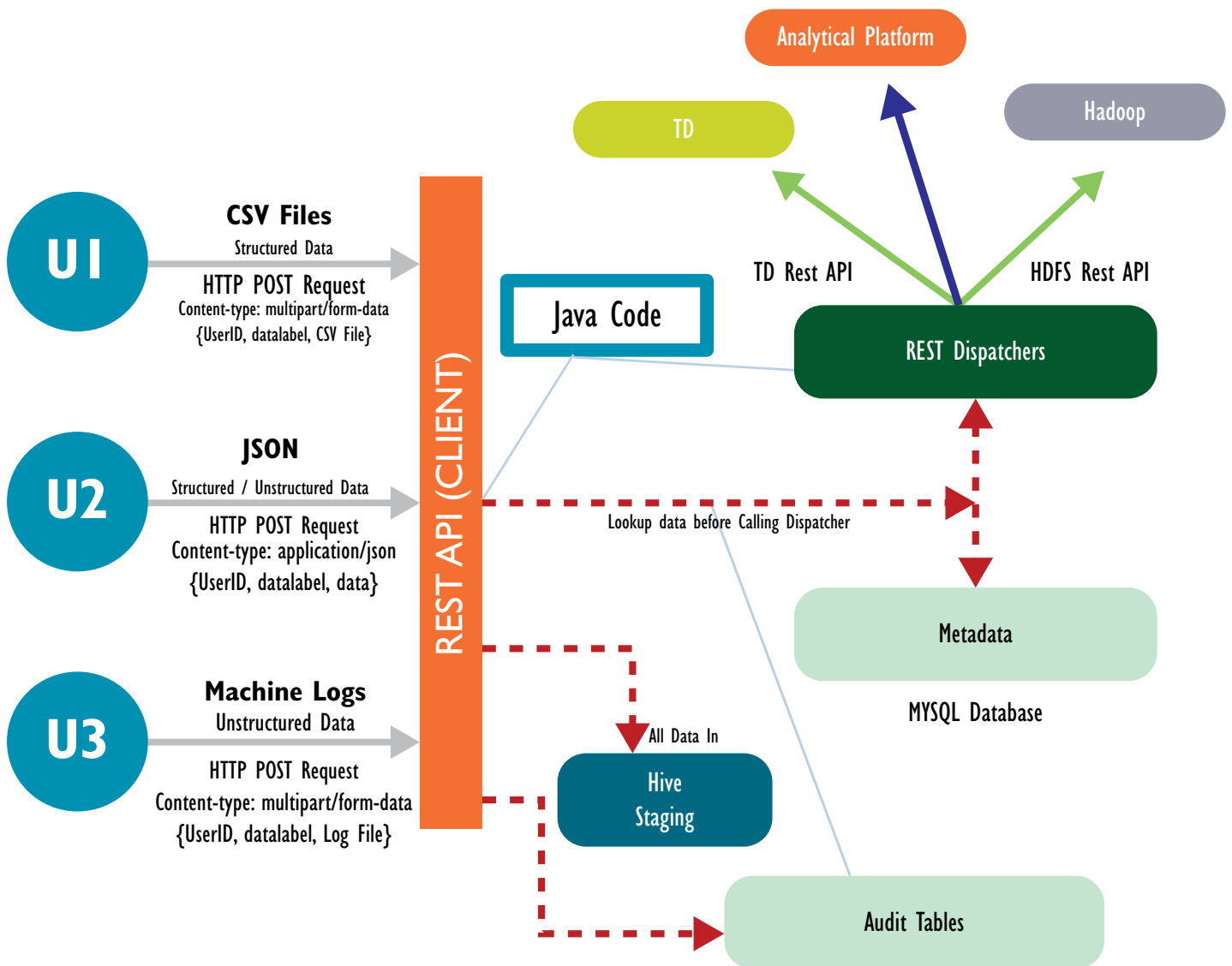
Data Ingestion

The architecture diagrams below elaborate how variety of data from different sources can be injected in to multiple data storage platforms (like UDA) with the help of REST framework.

Data Extraction

Data access from a multi-platform environment is easy with REST service as it provides the abstraction on top of the storage environment. Below is a sample architecture diagram on how REST acts as a Façade Layer for data storage.

Advantages of RESTFUL

End Users – Data Out

U1

JSON
Structured / Unstructured Data

HTTP GET Request
Re
Supporting 3 types:
/{user} : datalabels of user are displayed
/{user}/{datalabel}/metaData:
meta info of a datalabel
/{user}/{datalabel}*: data under datalabel
ColumnNames is an optional param

U2

CSV
Structured Data

HTTP GET Request
Request-Data: CSV
/{user}/{datalabel}*: data under datalabel
ColumnNames is an optional param

REST API (CLIENT)

Java Code

TD

Hadoop

REST Dispatchers

Lookup data before Calling Dispatcher

Metadata

MYSQL Database

All Data In

Audit Tables

Framework

There are many advantages for using REST in a multi-platform data warehouse environment:

- Being a public API, REST API is very easy to adopt and develop

- REST API for Teradata provides driverless connectivity to read and write data into Teradata database[v]. Similarly REST API for HDFS makes it easy to work with Hadoop Clusters

- Helps in work load balancing; no dependence on ETL tool or ESB (Enterprise Service Bus) for real-time integration

- REST works on top of HTTP; thus, only browser is needed for it to work

Conclusion

Real-time access to Hadoop along with other data warehouse platforms is promising because it provides the data pipeline not only for structured data but also to handle data types that the average data warehouse environment doesn't support. REST API's natural support to JSON objects add value, when new platforms like Mongo DB, Cassandra gets added in to the data warehouse environment.

The major benefit of using REST for real-time access is the low cost of development and the ease of deployment. Moreover REST naturally fits in the world of diverse data storage as it provides the perfect FAÇADE layer to inject and extract data from different platforms. Also, REST gives the ability for data scientists and business analysts to mix and m atch data on the fly without knowing where they reside. Also, now they don't have to wait for a day or two till the data load jobs complete in the conventional way.

Imagine moving machine data or web data in real-time using REST to Data Lake, processed using analytical platforms like Aster or in-memory analytical tools and stored in Teradata for business usage. At the end, the processed and report-ready data can be accessed using REST. All this is possible without the need to use highly priced business intelligence or ETL tools.

RESTFUL framework is indeed going to empower the internal customers and provide a cost effective way to integrate and access data in real-time.

References

1. http://www.datamation.com/applications/why-big-data-and-the-internet-of-things-are-a-perfect-match.html

2. http://tdwi.org/articles/2014/04/01/executive-summary-evolving-data-warehouse-architectures.aspx

3. http://javadevhell.blogspot.com/2010/11/rest-ful-web-service-basics-with.html

4. http://hortonworks.com/blog/webhdfs-%E2%80%93-http-rest-access-to-hdfs/

5. http://blogs.teradata.com/tdmo/rest-api-enables-driverless-connectivity/

About the Author

Pon Prabakaran Shanmugam is a Principal Consultant with Wipro Analytics practice. He possesses exhaustive data architecture experience in the Financial Industry with strong data modeling, integration and analytical skills, and is an enthusiastic & agile modeling proponent. He is also a strong believer of embracing open source technologies to make data architecture flexible and evolving.

About Wipro Ltd.

Wipro Ltd. (NYSE:WIT) is a leading Information Technology, Consulting and Business Process Services company that delivers solutions to enable its clients do business better. Wipro delivers winning business outcomes through its deep industry experience and a 360 degree view of "Business through Technology" - helping clients create successful and adaptive businesses. A company recognised globally for its comprehensive portfolio of services, a practitioner's approach to delivering innovation, and an organization wide commitment to sustainability, Wipro has a workforce of over 150,000, serving clients in 175+ cities across 6 continents.

For more information, please visit **www.wipro.com**

WIPRO

Applying Thought

DO BUSINESS BETTER