An Efficient HADOOP Frameworks SQOOP and Ambari for Big Data Processing

Mr. S. S. Aravinth
Assistant Professor
Department of Computer Science and Engineering
Knowledge Institute of Technology, Salem

Ms. A. Haseenah Begam
Department of Computer Science and Engineering
Knowledge Institute of Technology, Salem

Ms. S. Shanmugapriyaa
Department of Computer Science and Engineering
Knowledge Institute of Technology, Salem

Ms. S. Sowmya
Department of Computer Science and Engineering
Knowledge Institute of Technology, Salem

Mr. E. Arun
Department of Computer Science and Engineering
Knowledge Institute of Technology, Salem

Abstract

A Process having a large number of data affects the Operation. Due to this large augment of data, the industries are struggling to store, handle, and analyse the data. The normal data base systems are not enough to do the above mentioned activities.Then here comes the hadoop technology, In Hadoop the enormous data will be stored and processed effectively and efficiently. Hadoop is the technology which has many frameworks such as data integration, management, orchestration, monitoring, data serialization, data intelligence, storage, integration and access. So hadoop technology is used in which Sqoop tool is used ,it is a command-line interface application for transferring data between relational databases and hadoop . In hadoop scoop is the command line interface used for both Import and export from relational database to hadoop. In hadoop another tool called ambari is used. It is used to simplify Hadoop management processing of huge amount of data. It also works for provisioning, managing and monitoring of apache Hadoop clusters. In this paper the sqoop and ambari frameworks have been analysed with various parameters.

Keywords: Big Data, hadoop, Ambari , Sqoop, & Data processing

I. INTRODUCTION TO SQOOP

Apache Sqoop is a tool which is designed for efficient export or import of bulk data between Apache Hadoop and structured datastores. Presently Sqoop is a Top-Level Apache project which is a command line interface application written in java.

II. PURPOSE OF SQOOP

1) Sqoop is designed efficiently for the purpose of transferring huge amount of data between Apache hadoop and structured data stores such as relational.
2) It copies data quickly from external systems to hadoop.
3) It enables data imports from external data stores and enterprise data warehouses into hadoop.
4) It ensures fast performance by parallelizing data transfer and utilizes optimal system.
5) Sqoop supports analyses of data efficiently.
6) It even mitigates excessive loads to external systems.

III. WORKING OF SQOOP

1) Sqoop runs in hadoop cluster. It has access to hadoopcore.Sqoop use mappers to slice the incoming data.
2) Sqoop will communicate with the database store for fetching information called meta-data from relational datastore. This meta-data is being used for initiating java class by the Sqoop
3) Sqoop gets the metadata from DB store.
4) Sqoop will internally create a JAVA class using JDPC API. Sqoop will compile the java class using JDK and compare jar files.
5) Sqoop tries again to communicate with database store,once the jar files are created in order to find the split column which will enable Sqoop to fetch data from the database.

6) Finallysqoop will place the retrieved data into HDFS.

IV. SQOOP ARCHITECTURE

Sqoop have moved from Sqoop1 to Sqoop2 which means it has changed from client to server install. At this instance, Sqoop has web and command line access where client accesses Hbase and Hive. Oozie uses REST API. Client can run in two modes- interactive mode and batch mode.

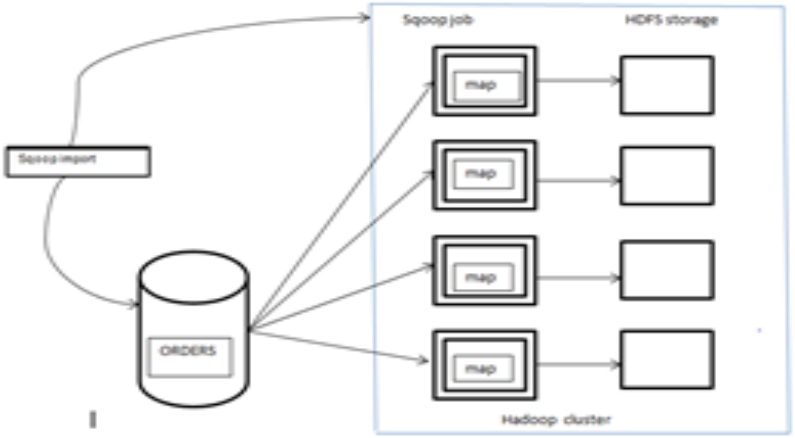

Fig. 1: SQOOP Architecture

V. SQOOP – IMPORTS

A. Import:

This is a sub-command which instructs Sqoop to initiate import.

To import data, Sqoop examines the database for collecting the necessary metadata.

Sqoop submits to cluster- A Map-only Hadoop job then the Map-only job performs the data transfer using the metadata.

After the data being imported,it is saved in a directory HDFS(Hadoop Distributed File System) or even any alternative directory can be mentioned by the user for where the imported data is to be stored.

These files contain comma delimited fields by default, in which there are new lines separating various records. Here also the user can override the format in which the data is copied if the field separator and the record terminator characters are mentioned explicitly.

Sqoop still enables different data formats for data imports. For instance, Avro data format can be used by the user for easily importing data. Example: avrodatafile with the import command. To perform import operation, Sqoop provides several options too.

Examples for importing data into Hbase:

1) Hbase-create-table: It instructs Sqoop to create the Hbase table.
2) Hbase-table: It specifies table name to use.

VI. SQOOP – EXPORTS

A. Export:

This is a sub-command which instructs Sqoop to initiate export.

To export data, Sqoop examines the database for collecting the necessary metadata. It is followed by transferring those data.

This input dataset is divided into a number of splits using Sqoop. Then Sqoop pushes these splits into database using individual map tasks.

Each of these map task performs this transfer over many transactions for ensuring minimal utilization and optimal throughput.

VII. SQOOP CONNECTORS

Sqoop connects with external systems which have optimized import and export facitilies, or do not support native JDBC.

Connectors are plugin components based on Sqoop's Extension framework and can use it to any extistingSqoopinstallion. Once a connector is installed, Sqoop can use it efficiently for transfering data between Hadoop and external database stores supported by the connectors.

VIII. APPLICATIONS OF SQOOP

1) Sqoop is used to exchange data between Hadoop and IBM Netezza data warehouse appliance by Online Marketer Coupons.com.
2) The Apollo group, Education Company uses Sqoop to extract data from databases and to inject the results into relational database from hadoop jobs.
3) In addition to this, there are countless other hadoop users who use Sqoop to efficiently transfer their data.

A. Apache Ambari:

Apache ambariis a completely open operational framework for provisioning, managing and monitoring apache hadoopcluster.Ambari includes an intuitive collection of operator tools and Aset of APIs that hidden the difficult task of hadoop,it contract the operation of cluster. With hundreds of years combined experience of Hortonwork along with the hadoop community have answered the call to deliver the key services required for enterprice hadoop.it does ambari enables the system administrators to manage, monitor and to provision a hadoop cluster and it is also used to integrate hadoop with infrastructure of the existing enterprise.

- It simplified operations like ability to perform automated stack upgrades
- It enterprise readiness like add support for the windows platforms
- It is usability enhancements improvements to alerts and metrics collection

B. Manage and Monitor Hadoop Improvements:

Our open approach via apache ambari and we are excited to have HP, pivotal and VMware jump on word to support ambari with some of the others in the data center like Teradata and Microsoft. This open ness allows everyone to like and enjoy the new features as they are delivered and the community of ambari is developing in such an amazing rate in HDP 2.2,

C. Configuration Versioning and History:

A comprehensive approach is delivering in order to address configuration.

D. Extent Ambari with Custom Views:

Framework here is a systematic way. When third parties are allowed to plug in new resource which is developed in ambari container.

E. Ambari Blue Prints Deliver a Template Approach To Cluster Deployment:

Ambari blueprints are declarative definition of cluster and specifying the stack then the component and configurations to materialized a hadoop cluster instance

F. Recent Ambari Releases:

1.6.0: Introduced ambari blueprints for automating cluster installs,Improved usability guardrails with more host and environment checks,Support for postgreSQL database

1.5.0. rollingrestarts,Service and host maintenance mode,Bulk host operation.

G. Provision a Hadoop Cluster:

No matter the size of your hadoop cluster,the development and maintenance of hosts is contracted using ambari. It include web interface that allow you to easily provision,configure and test all the hadoop service. It also provides powerful ambari blueprints API for automatic cluster installations without user interventions

H. Manage a Hadoop Cluster:

Ambari provides amazing tools to managing the cluster in simplified way. The web interface allows the user to control the lifecycle of hadoopcomponents and services and it modifies the configurations and manages the ongoing growth of the user's cluster.

I. Integrated Hadoop with The Enterprise

Ambari provides you a restful API which enables integration with already existing tools, such as Teradata viewpoint and Microsoft system center, which is used to merge your established operational processes with hadoop.
The important features are:

- Wizard-driven interface: Hadoop can be installed in many hosts.
- API-driven installations:ambari blueprints for automated provisioning.
- Granular control: precise management of hadoop services and component lifecycles.
- Configuration histories: ongoing management of hadoop service configurations.
- Extensible framework: brings custom services under management via ambari stacks.
- Usability improvements: innovative user experience via ambari views.
- RESTful APIs:enables integration with enterprise system.

J. Hortonworks Collaboration for Ambari:

Hortonworks is the open source marketing. It allows providing the product management guidance for hadoop to mainstream enterprises in orgin of hadoop.

IX. CONCLUSION

This paper makes a research and gives the brief and elaborative knowledge on, An Efficient Hadoop Technology Frameworks Sqoop and Ambari for Big Data Analysis and Processing.This paper contains all the basics and technical things related to sqoop and ambari. and also it gives a clear view on sqoop and ambari for people who really want to know hadoop technology frame work.

REFERENCES

[1] http://hortonworks.com/hadoop/ambari/
[2] http://link.springer.com/content/pdf/10.1007%252F978-1-4302-4864-4_20.pdf
[3] http://www.tutorialspoint.com/sqoop/sqoop_pdf_version.htm