EFFECTIVE QUERIES OF PIG IN OOZIE AUTOMATION

[1]S.Sivasankara Rao, [2] Padamarajani, [3] R.Narender
[1]Associate Professor, [2] Assistant Professor, [3] Assistant Professor
Department of CSE,
Guru Nanak Institutions Technical Campus,
Ibrahimpatnam,Telangana,India-501506.

ABSTRACT

Oozie is widely used in several large production clusters across major enterprises to schedule Hadoop jobs. It simplifies the process of creating workflows and managing coordination among jobs. Oozie offers the ability to combine multiple jobs sequentially into one logical unit of work. An Oozie workflow is a collection of actions arranged in a directed acyclic graph (DAG).Pig is an abstraction over Map Reduce. It is a tool/platform which is used to analyse larger sets of data representing them as data flows. Pig is generally used with Hadoop; we can perform all data manipulation operations in Hadoop using Pig.

The purpose of this is to provide automation through Oozie where large sets of data are analysed using Pig queries thereby reducing the time complexity.

Keywords: Hadoop, Apache Oozie, Map Reduce, Apache Pig.

1. INTRODUCTION

Automation here stands for implementing, i.e. a process or procedure is performed without human assistance or humans typing the code directly in this case. Oozie is a workflow scheduler system to manage Apache Hadoop jobs. It was clear that there was a need for a general-purpose system to run multistage.

Hadoop jobs with the requirements like, it should use an adequate and well-understood programming model to facilitate, its adaption and to reduce developer ramp-up time. It should be easy to troubleshoot and recover jobs when something goes wrong. It should be extensible to support new types of jobs. It should scale to support several thousand concurrent jobs. Jobs should run in a server to increase reliability. It should be a multitenant service to reduce the cost of operation. The Benefits of Oozie are

1. Oozie Workflow jobs are Directed A cyclical Graphs (DAGs) of actions[11].
2. Oozie is designed to scale in a Hadoop cluster. Each job will be launched from a different data node. This means that the workflow load will be balanced and no single machine will become overburdened by launching workflows.
3. Oozie is well integrated with Hadoop security.Oozie knows which user submitted the job and will launch all actions as that user, with the proper privileges. It will handle all the authentication details for the user as well.
4. Oozie is still recommended workflow scheduler due to its ability to handle complexity, ease of integration with established and emerging Hadoop components , and the growing ecosystem of projects, such as Apache Falcon, that rely on its workflow. Oozie also remains one of the more challenging schedulers to learn and master.
5. Apache Oozie remains the most sophisticated and powerful workflow scheduler for managing Apache Hadoop jobs.
6. Oozie allows you to start from failure Oozie to restart a job from a specific node in the graph or to skip specified failed nodes.
7. A vulnerability in the Oozie Server allows a cluster user to read private files owned by the user running the Oozie Server process.
8. Oozie workflows can be scheduled using coordinators. Coordinators can have start time, end time, and frequency parameters. All those have an effect on when and how often coordinator actions (occurrences) of the same coordinator job (definition) will schedule to run.

2. PROPOSED SYSTEM

- Automate data load process into HDFS, HIVE[2]
- Suppose we are receiving daily feeds from any source (database, log files or anything) and need to make it available in HDFS and in turn in Pig for analytics. A simple solution may be using Oozie. Here, we schedule the process to run on every Business Day at a specific time. Then Following features are included:-

1. Moving the data from RDBMS to HDFS (Using Spool)
2. Running Map-Reduce
3. Storing the output in HDFS/HIVE
4. Storing the output to RDBMS (Using Spool)

Big data analytics define the analysis of large amount of data to get the useful information and uncover the hidden patterns. Real Time Literature Review about the Big data. According to 2013, Facebook has 1.11 billion people active accounts from which 751 million using Facebook from a mobile.

Describe the big data content, its scope, methods, samples, advantages and challenges of Data. The critical issue about the Big data is the privacy and security. Big data samples describe the review about the atmosphere, biological science and research. Life sciences etc . By this paper, we can conclude that any organization in any industry having big data can take the benefit from its careful analysis for the problem solving purpose. Using Knowledge Discovery from the Big data easy to get the information from the complicated data sets. The overall Evaluation describe that the data is increasing and becoming complex.

[3]Oozie is an Open Source Java Web-Application available under Apache license 2.0. It is responsible for triggering the workflow actions, which in turn uses the hadoop execution engine to actually execute the task. Hence, Oozie is able to leverage the existing Hadoop machinery for load balancing, fail-over, etc.

Oozie detects completion of tasks through call back and polling. When Oozie starts a task, it provides a unique call-back HTTP URL to the task, and notifies that URL when it is complete. If the task fails to invoke the call back URL, Oozie can poll the task for completion.

3. SYSTEM ARCHITECTURE

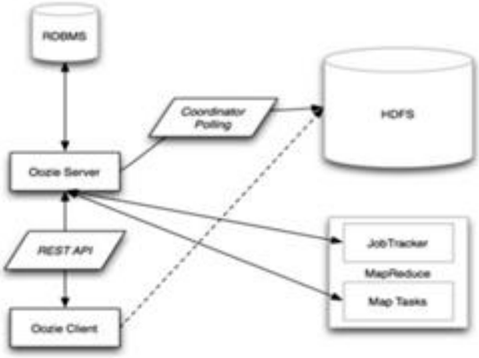

Fig 1:Oozie Architecture

Oozie is used to schedule one or more Hadoop jobs on a regular basis. It is mostly used in production environment to schedule recurring jobs. Oozie is written in command lines around jobs or a bundle of jobs. There is a control dependency - which means one job cannot run until the first job is completed. The workflow can start jobs in a remote system and the system sends a notification to the Oozie server - once the job is completed.

- Oozie has control flow nodes and action nodes. Control flow nodes control the start and end of the workflow and also manages the execution in the workflow and action nodes trigger the jobs in the scheduler. Action nodes specify the type of action that needs to be performed - MapReduce jobs or Scripts.
- Oozie was developed as an alternative to manual and ad-hoc approaches to shell scripts, job control that were there to schedule jobs in the workflow. Oozie detects the completion of jobs by two actions call-back and polling.

There are three basic types of Oozie jobs[]:

- **Oozie Workflow** jobs are Directed ACyclical Graphs (DAGs), specifying a sequence of actions to execute. the Workflow job has to wait[2].
- **Oozie Coordinator** jobs are recurrent Oozie Workflow jobs that are triggered by time and data availability.
- **Oozie Bundle** provides a way to package multiple coordinator and workflow jobs and to manage the lifecycle of those jobs[3].

4.PIG ARCHITECTURE

Applications of Apache Pig: Apache Pig is generally used by data scientists for performing tasks involving ad-hoc processing and quick prototyping. Apache Pig is used –

- To process huge data sources such as web logs.
- To perform data processing for search platforms.
- To process time sensitive data loads.

The language used to analyze data in Hadoop using Pig is known as Pig Latin.

To perform a particular task Programmers using Pig, programmers need to write a Pig script using the Pig Latin language, and execute them using any of the execution mechanisms (Grunt Shell, UDFs, Embedded). After execution, these scripts will go through a series of transformations applied by the Pig Framework, to produce the desired output.

[1]Apache Pig is a data flow language. It is a high level language. Performing a Join operation in Apache Pig is pretty simple. Any novice programmer with a basic knowledge of SQL can work conveniently with Apache Pig. Apache Pig uses multi-query approach, thereby reducing the length of the codes to a great extent. There is no need for compilation. On execution, every Apache Pig operator is converted internally into a MapReduce job.

MapReduce is a data processing paradigm. MapReduce is low level and rigid. It is quite difficult in MapReduce to perform a Join operation between datasets. Exposure to Java is must to work with MapReduce. MapReduce will require almost 20 times more the number of lines to perform the same task. MapReduce jobs have a long compilation process.

Internally, Apache Pig converts these scripts into a series of MapReduce jobs, and thus, it makes the programmer's job easy. The architecture of Apache Pig is shown below.

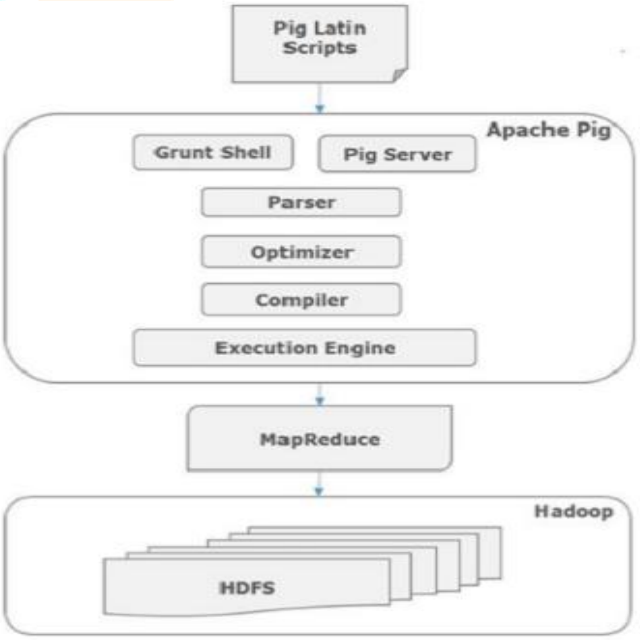

Fig 2:Architecture of Apache Pig

4.1 Apache Pig Components

As shown in the figure, there are various components in the Apache Pig framework. Let us take a look at the major components[3].

4.1.1.Parser

Initially the Pig Scripts are handled by the Parser. It checks the syntax of the script, does type checking, and other miscellaneous checks. The output of the parser will be a DAG (directed acyclic graph), which represents the Pig Latin statements and logical operators.

In the DAG, the logical operators of the script are represented as the nodes and the data flows are represented as edges.

4.1.2.Optimizer

The logical plan (DAG) is passed to the logical optimizer, which carries out the logical optimizations such as projection and pushdown.

4.1.3Compiler

The compiler compiles the optimized logical plan into a series of MapReduce jobs.

4.1.4.Execution engine

Finally the MapReduce jobs are submitted to Hadoop in a sorted order. Finally, these MapReduce jobs are executed on Hadoop producing the desired results.

[5]A good example of a Pig application is the ETL transaction model that describes how a process will extract data from a source, transform it according to a rule set and then load it into a data store. Pig can ingest data from files, streams or other sources using the User Defined Functions (UDF). Once it has the data it can perform select, iteration, and other transforms over the data. Again the UDF feature allows passing the data to more complex algorithms for the transform. Finally Pig can store the results into the Hadoop Data File System.

[1]Apache Pig is a platform for analyzing large data sets that consists of a high-level language for expressing data analysis programs, coupled with infrastructure for evaluating these programs. The salient property of Pig programs is that their structure is amenable to substantial parallelization, which in turns enables them to handle very large data sets. At the present time, Pig's infrastructure layer consists of a compiler that produces sequences of Map-Reduce programs, for which large-scale parallel implementations already exists . Pig's language layer currently consists of a textual language called Pig Latin. It is trivial to achieve parallel execution of simple, "embarrassingly parallel" data analysis tasks. Complex tasks comprised of multiple interrelated data transformations are explicitly encoded as data flow sequences, making them easy to write, understand, and maintain.

To implement Pig under Oozie with the help of Hadoop it create 3 files for any Project that’s required for Oozie to run Pig.

[4]Job properties file Sets properties or arranges values to the parameters present in the workflow file. WorkFlow.xml file defines the structure of the program with the main modules and agents for which the path is set by Job properties file. This file helps Oozie to define the path where to go next or what's in the queue next. The id.pig: This file has the Pig queries that need to be executed.

- We need to start 3 terminals that’s in each we should start Oozie, Hadoop and Hive Server Respectively.

CONCLUSION

- We learnt how to schedule the Pig job using Oozie. In production, where we need to run the same job for multiple times, or, we have multiple jobs that should be executed one after another, we need to schedule the job using some scheduler. There are multiple ways to automate jobs, however, here we worked with Oozie. We begun with understanding what Oozie is and Oozie job scheduling is.
- Oozie, an open source Apache project is a job scheduler that manages Hadoop jobs. In short, Oozie schedules long list of works sequentially into one job.
- To schedule Pig job using Oozie, you need to write a Pig-action. The Oozie job will consist of mainly three things. They are workflow ,job properties and Hive Script.

By this paper, we can conclude that any organization in any industry having big data can take the benefit from its careful analysis for the problem solving purpose. Using Knowledge Discovery from the Big data easy to get the information from the complicated data sets. The overall Evaluation describe that the data is increasing and becoming complex.

FUTURE ENHANCEMENT

How to manage data is the key concern for all the commercial houses for which these rely on Bigdata solutions, therefore increasing its scope manifold. However, with more and more business houses depending on Bigdata Solutions for enterprise success, there is an ever-rising demand for efficient Bigdata professionals, who have an expertise in the areas of Hadoop and related technologies such as Pig, Hive, Spool, Kafka, and Oozie.

REFERENCES

[1].Apache Oozie The Work Flow Scheduler for Hadoop By Mohammad Kamrul Islam & Aravind Srinivasan.

[2].Bakshi,K.,(2012)”Considerations for Bigdata:Architecture and approach”.

[3].Mukherjee,A;Datta,J.;Jorapur,R.;Haloi,S.;Akram,W.,(18-22Dec.,2012),“shared disk big data analytics with Apache Hadoop”.

[4]. Distributed Log Collection for Hadoop by Steve Hoffman.

[5].Jonathan Stuart Ward and Adam Barker “**Undefined By Data: A Survey of Big Data Definitions”** Stamford, CT: Gartner, 2012.

[6].Ms.Vibhavari Chavan, Prof. Rajesh. N. Phursule. “*Survey paper on Big Data*”, International Journal of Computer Science and Information Technologies, Vol. 5 (6) , 2014, 7932-7939.

[7]. Sabia, Love Arora,” Technologies to Handle Big Data: A Survey”, International Conference on Communication, Computing & Systems (ICCCS–2014).

[8].Harshawardhan S. Bhosle , Prof. Devendra P. Gadekar, “ A Review Paper on Big Data and Hadoop“, International Journal of Scientific and Research Publications, Volume 4, Issue 10, October 2014, ISSN 2250-3153.

[9].Definitions of Big Data - open tracker, http://www.pcmag.com/encyclopedia/term/62849/big-data.

[10].Rathod, Chauhan, "A Survey on Big Data Analysis Techniques" IJSRD - International Journal for Scientific Research & Development Vol. 1, Issue 9, 2013 ISSN (online): 2321-0613.

[11].Apache Oozie available at: https://oozie.apache.org.