On-Demand Analytics Engine

Aditya Joshi, Lokesh Jain, Rishabh Srivastava, Syeda Afnan

On Demand Analytics Engine is a SaaS service which is capable of executing Data Analysis jobs on various Hadoop related platforms

Problem Statement

- Develop a system for running jobs on frameworks like Hadoop, Mahout, Giraph and RHadoop which would otherwise be an herculean task
- An engine which could directly interact providing APIs to the end-user, data scientists, web-developers and software programmers

Approach

- Frameworks were studied and selected according to purpose
- Most used algorithms for these frameworks were selected
- Abstraction layer was designed in accordance with the requirements of the algorithms as well as framework
- Control Flow for the application was designed and created.
- REST/JSON links were created

Component Design

The tool is a JAVA application that performs the following tasks:

Tomcat server receives the query from the REST API

↓

The GET string is parsed to select the API that user wants to execute.

↓

Related methods are executed which in turn execute the algorithms on user data

↓

Corresponding JSON response is generated.

Challenges

PrivilegedException : Namespace of the DFS in the slave node mismatches with the DFS of the master node.

HostNameException : The master may not be able to resolve the host name of other nodes

RPackageIssue : rmr package version available at the current Github repository was not compatible with RHadoop

Abstraction Layer

REST (Input) and JSON (Output)

Java Core Layer

Takes input in REST format and responds in JSON format after processing in JSON format.

Mahout

Build scalable machine learning libraries

RHadoop

R interface to Hadoop

Giraph

Graph processing platform for large-scale algorithms

Hadoop

Open-source framework that supports data-intensive distributed applications

Cluster Layer

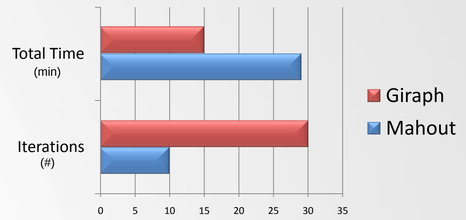
Referred as HDFS, the layer is capable of storing and processing large amounts of data. It is a fault-tolerant, scalable, distributed file system built on commodity hardware.

Applications

- It allows the organizations to have an easy-to-use Data Analytics framework which can be used through the provided web application or custom-built applications with JSON capability.
- In business deployment, the users can be charged on either month-to-month basis or data/bandwidth usage.

Evaluation

Setup* analyzed 6 Million vertices from Wikipedia



Documentation

The complete documentation of the analytics engine is given in the Documentation folder. This includes:

- Algorithm :
  Explanation of execution and addition of algorithms
- API Specs :
  Explanation of REST APIs to be used to interact with the system from the front-end
- Exceptions :
  Encountered errors and exceptions with their possible solutions

API Spec.

1. Execute
   1. Execute Algorithm
   2. Execute Script File
2. Add
   1. Add Framework
   2. Add Algorithm
3. Remove
   1. Remove Framework
   2. Remove Algorithm
4. Status
5. List
   1. List Frameworks
   2. List Algorithms
   3. List Algorithms files

Features

- Extensible, portable, usable
- Dataset mounted instead of copying.

Future Work

- **Yarn :** We can also upgrade to Yarn, which is specifically built for data processing
- **Scheduling :** Improved scheduling algorithms to manage multiple jobs running in parallel
- **Performance :** Select the Framework for user's task using Machine Learning approaches

Limitations

- The file system used might not be able to extend for other file systems like GFS.
- The progress percentage of the running jobs can't be provided. Only the completion status is given.

References

- Commercial project "ondemandanalysis.com" and Cloudera "cloudera.com"
- Performance Stats referred from slideshare.net/sscdotopen/large-scale

Acknowledgement

Prof. Vasudeva Varma, Mr. Reddy Raja, Mr. Janakiraman
Mr. Dharmesh Kakadia, Mr. Ankit Patil and other TAs

Github : https://github.com/ljain/cloud-analytics