Apache Tajo

tutorialspoint

SIMPLY EASY LEARNING

www.tutorialspoint.com

https://www.facebook.com/tutorialspointindia https://twitter.com/tutorialspoint

About the Tutorial

Apache Tajo is an open-source distributed data warehouse framework for Hadoop. Tajo was initially started by Gruter, a Hadoop-based infrastructure company in south Korea. Later, experts from Intel, Etsy, NASA, Cloudera, Hortonworks also contributed to the project.

Tajo refers to an ostrich in Korean language. In the year March 2014, Tajo was granted a top-level open source Apache project. This tutorial will explore the basics of Tajo and moving on, it will explain cluster setup, Tajo shell, SQL queries, integration with other big data technologies and finally conclude with some examples.

Audience

Before proceeding with this tutorial, you must have a sound knowledge on core Java, any of the Linux OS, and DBMS.

Prerequisites

This tutorial has been prepared for professionals aspiring to make a career in big data analytics. This tutorial will give you enough understanding on Apache Tajo.

Disclaimer & Copyright

All the content and graphics published in this e-book are the property of Tutorials Point (I) Pvt. Ltd. The user of this e-book is prohibited to reuse, retain, copy, distribute or republish any contents or a part of contents of this e-book in any manner without written consent of the publisher.

We strive to update the contents of our website and tutorials as timely and as precisely as possible, however, the contents may contain inaccuracies or errors. Tutorials Point (I) Pvt. Ltd. provides no guarantee regarding the accuracy, timeliness or completeness of our website or its contents including this tutorial. If you discover any errors on our website or in this tutorial, please notify us at contact@tutorialspoint.com.

Table of Contents

About the Tutorial i
Audience i
Prerequisites i
Disclaimer & Copyright i
Table of Contents ii

1. Apache Tajo — Introduction 1
Distributed Data Warehouse System 1
Overview of SQL on Hadoop 1
What is Apache Tajo 1
Features of Apache Tajo 2
Benefits of Apache Tajo 2
Use Cases of Apache Tajo 2
Storage and Data Formats 3

2. Apache Tajo — Architecture 4
Workflow 5

3. Apache Tajo — Installation 6
Verifying Java installation 6
Download JDK 6
Verifying Hadoop Installation 6
Apache Tajo Installation 7
Extract Tar File 7
Set Environment Variable 7
Start Tajo Server 7
Launch Tajo Shell (Tsql) 8
Quit Tajo Shell 8
Web UI 9
Stop Tajo 9

4. Apache Tajo — Configuration Settings 10
Basic Settings 10
Distributed Mode Configuration 10
Master Node Configuration 11
Catalog Configuration 11
Worker Configuration 12

5. Apache Tajo — Shell Commands 14
Meta Commands 14
Admin Commands 17
Session Variables 19

6. Apache Tajo — Data Types 21

7. Apache Tajo – Operators 23
Arithmetic Operators 23
Relational Operators 24
Logical Operators 25
String Operator 26
Range Operator 27

8. Apache Tajo – SQL Functions 28

9. Apache Tajo – Math Functions 29
abs(x) 30
cbrt(x) 31
ceil(x) 31
floor(x) 31
pi() 32
radians(x) 32
degrees(x) 33
pow(x,y) 33
div(x,y) 33
exp(x) 34
sqrt(x) 34
sign(x) 35
mod(n,m) 35
round(x) 35
cos(x) 36
asin(x) 36
acos(x) 37
atan(x) 37
atan2(y/x) 37
Data Type Functions 38
to_bin(x) 38
to_char(int,text) 39
to_hex(x) 39

10. Apache Tajo – String Functions 40
concat (string1, ..., stringN) 41
length (string) 41
lower (string) 42
upper (string) 42
asci (char) 43
bit_length (string) 43
char_length (string) 43
octet_length (string text) 44
digest (string, method) 44
initcap (string) 45
md5 (string) 45
left (string, size) 45
right (string, size) 46
locate (string, source text, target text, start_index) 46

strposb (source, target) ... 47
substr (string, start, length) ... 47
trim(string,char) ... 47
split_part (string, delimiter, index) ... 48
regexp_replace (string text, pattern text, replacement text) ... 48
reverse(string) ... 49

11. Apache Tajo – DateTime Functions ... 50
current_date() ... 51
add_days(date,day) ... 52
add_months(date,month) ... 52
current_time() ... 53
extract(century) ... 53
extract(day) ... 54
extract(decade) ... 54
extract (day of week) ... 54
extract (day of year) ... 55
extract(hour) ... 55
extract (isodow) ... 56
extract (isoyear) ... 56
extract (microseconds) ... 56
extract (millennium) ... 57
extract (milliseconds) ... 57
extract (minute) ... 58
extract (quarter) ... 58
date_part (month,date) ... 58
now() ... 59
to_char (timestamp,format) ... 59
to_date (date, format) ... 60
to_timestamp (date,format) ... 60

12. Apache Pajo – JSON Functions ... 61
json_extract_path_text (json text, json_path text) ... 61
json_array_get(json_array text, index int4) ... 62
json_array_contains (json_array text, value any) ... 62
json_array_length(json_array text) ... 62

13. Apache Tajo – Database Creation ... 64
Create Database Statement ... 64
Show Current Database ... 64
Connect to Database ... 64
Drop Database ... 65

14. Apache Tajo — Table Management ... 66
External Table ... 66
Table Properties ... 66
Managed Table ... 67
Tablespace ... 67
Tablespace Creation ... 68
Configure Tablespace ... 69
Data formats ... 69

JSON 70
Parquet 71
RCFile 71
SequenceFile 72
ORC 72

15. Apache Tajo — SQL Statements 73
Create Table Statement 73
Create Database 73
Show Table 74
List table 75
Insert Table Statement 76
Insert Records 76
Fetch records 77
Add Column 78
Set Property 78
Select Statement 79
Where Clause 80
Distinct Clause 81
Group By Clause 81
Having Clause 82
Order By Clause 83
Create Index Statement 83
Drop Table Statement 84

16. Apache Tajo – Aggregate & Window Functions 86
Aggregation Functions 86
Avg(exp) 86
corr (exp1,exp2) 87
count() 87
max (exp) 88
min(exp) 88
sum(exp) 88
last_value(exp) 89
Window Function 89
rank() 90
row_number() 91
lead (value, offset, default) 91
lag (value, offset, default) 92
first_value (value) 93
last_value(value) 93

17. Apache Tajo — SQL Queries 95
Predicates 95
Like Predicate 96
Using NULL Value in Search Conditions 97
Explain 99
Joins 101
Inner Join 102
Left Outer Join 103
Right Outer Join 104

Full Outer Join 104
Cross Join 105
Natural Join 106
Self Join 107

18. Apache Tajo — Storage Plugins 109
storage-site.json 109
PostgreSQL Storage Handler 109

19. Apache Tajo — Integration with HBase 111
Set Environment Variable 111
Create an External Table 111
Create Table in HBase 112

20. Apache Tajo — Integration with Hive 114
Set Environment Variable 114
Catalog Configuration 114

21. Apache Tajo — OpenStack Swift Integration 115
Core-site.xml 115
conf/tajo-env.h 115
Create Table 116

22. Apache Tajo — JDBC Interface 117
Download JDBC Driver 117
Set Class Path 117
Connect to Tajo 117
Java Application 118

23. Apache Tajo — Custom Functions 121

1. APACHE TAJO — INTRODUCTION

Distributed Data Warehouse System

Data warehouse is a relational database that is designed for query and analysis rather than for transaction processing. It is a subject-oriented, integrated, time-variant, and non-volatile collection of data. This data helps analysts to take informed decisions in an organization but relational data volumes are increased day by day.

To overcome the challenges, distributed data warehouse system shares data across multiple data repositories for the purpose of Online Analytical Processing(OLAP). Each data warehouse may belong to one or more organizations. It performs load balancing and scalability. Metadata is replicated and centrally distributed.

Apache Tajo is a distributed data warehouse system which uses Hadoop Distributed File System (HDFS) as the storage layer and has its own query execution engine instead of MapReduce framework.

Overview of SQL on Hadoop

Hadoop is an open-source framework that allows to store and process big data in a distributed environment. It is extremely fast and powerful. However, Hadoop has limited querying capabilities so its performance can be made even better with the help of SQL on Hadoop. This allows users to interact with Hadoop through easy SQL commands.

Some of the examples of SQL on Hadoop applications are Hive, Impala, Drill, Presto, Spark, HAWQ and Apache Tajo.

What is Apache Tajo

Apache Tajo is a relational and distributed data processing framework. It is designed for low latency and scalable ad-hoc query analysis.

- Tajo supports standard SQL and various data formats. Most of the Tajo queries can be executed without any modification.

- Tajo has **fault-tolerance** through a restart mechanism for failed tasks and extensible query rewrite engine.

- Tajo performs the necessary **ETL (Extract Transform and Load process)** operations to summarize large datasets stored on HDFS. It is an alternative choice to Hive/Pig.

The latest version of Tajo has greater connectivity to Java programs and third-party databases such as Oracle and PostGreSQL.

Features of Apache Tajo

Apache Tajo has the following features:

- Superior scalability and optimized performance
- Low latency
- User-defined functions
- Row/columnar storage processing framework.
- Compatibility with HiveQL and Hive MetaStore
- Simple data flow and easy maintenance.

Benefits of Apache Tajo

Apache Tajo offers the following benefits:

- Easy to use
- Simplified architecture
- Cost-based query optimization
- Vectorized query execution plan
- Fast delivery
- Simple I/O mechanism and supports various type of storage.
- Fault tolerance

Use Cases of Apache Tajo

The following are some of the use cases of Apache Tajo:

Data warehousing and analysis

Korea's SK Telecom firm ran Tajo against 1.7 terabytes worth of data and found it could complete queries with greater speed than either Hive or Impala.

Data discovery

The Korean music streaming service Melon uses Tajo for analytical processing. Tajo executes ETL (extract-transform-load process) jobs 1.5 to 10 times faster than Hive.

Log analysis

Bluehole Studio, a Korean based company developed TERA — a fantasy multiplayer online game. The company uses Tajo for game log analysis and finding principal causes of service quality interrupts.

Storage and Data Formats

Apache Tajo supports the following data formats:

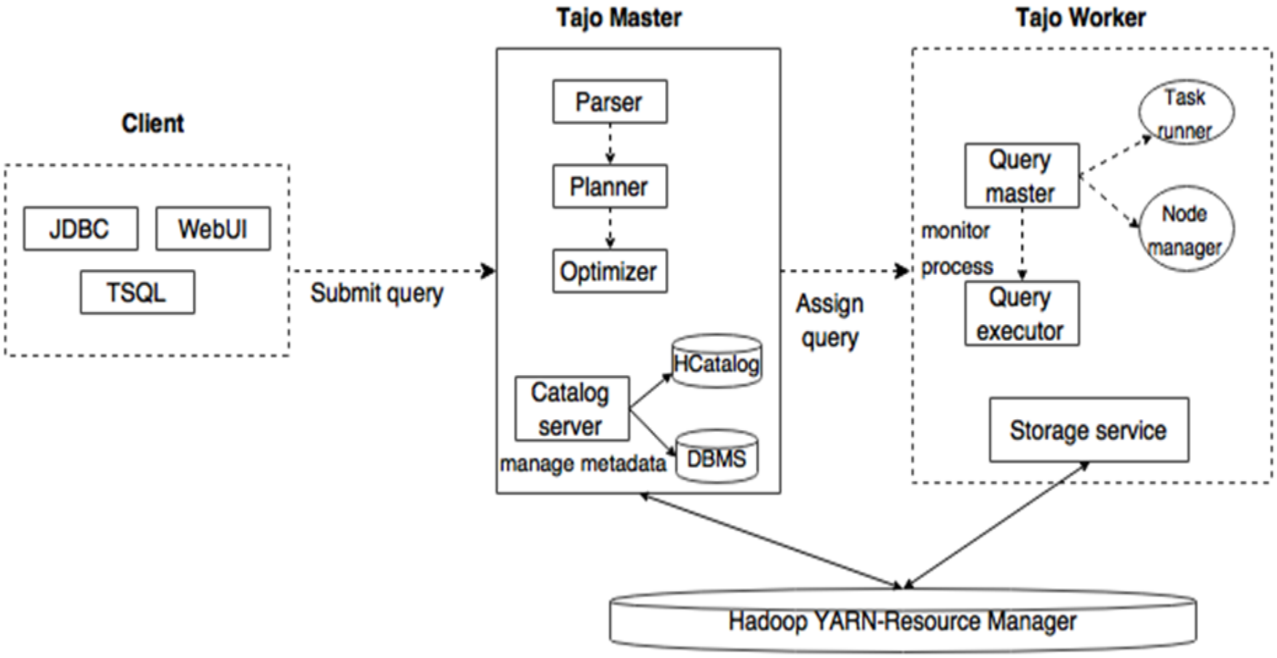
- JSON
- Text file(CSV)
- Parquet
- Sequence File
- AVRO
- Protocol Buffer
- Apache Orc

Tajo supports the following storage formats:

- HDFS
- JDBC
- Amazon S3
- Apache HBase
- Elasticsearch

2. APACHE TAJO — ARCHITECTURE

The following illustration depicts the architecture of Apache Tajo.



The following table describes each of the components in detail.

Component	Description
Client	**Client** submits the SQL statements to the Tajo Master to get the result.
Master	Master is the main daemon. It is responsible for query planning and is the coordinator for workers.
Catalog server	Maintains the table and index descriptions. It is embedded in the Master daemon. The catalog server uses Apache Derby as the storage layer and connects via JDBC client.

Worker	Master node assigns task to worker nodes. TajoWorker processes data. As the number of TajoWorkers increases, the processing capacity also increases linearly.
Query Master	Tajo master assigns query to the Query Master. The Query Master is responsible for controlling a distributed execution plan. It launches the TaskRunner and schedules tasks to TaskRunner. The main role of the Query Master is to monitor the running tasks and report them to the Master node.
Node Managers	Manages the resource of the worker node. It decides on allocating requests to the node.
TaskRunner	Acts as a local query execution engine. It is used to run and monitor query process. The TaskRunner processes one task at a time. It has the following three main attributes: • Logical plan - An execution block which created the task. • A fragment - an input path, an offset range, and schema. • Fetches URIs
Query Executor	It is used to execute a query.
Storage service	Connects the underlying data storage to Tajo.

Workflow

Tajo uses Hadoop Distributed File System (HDFS) as the storage layer and has its own query execution engine instead of the MapReduce framework. A Tajo cluster consists of one master node and a number of workers across cluster nodes.

The master is mainly responsible for query planning and the coordinator for workers. The master divides a query into small tasks and assigns to workers. Each worker has a local query engine that executes a directed acyclic graph of physical operators.

In addition, Tajo can control distributed data flow more flexible than that of MapReduce and supports indexing techniques.

The web-based interface of Tajo has the following capabilities:

- Option to find how the submitted queries are planned