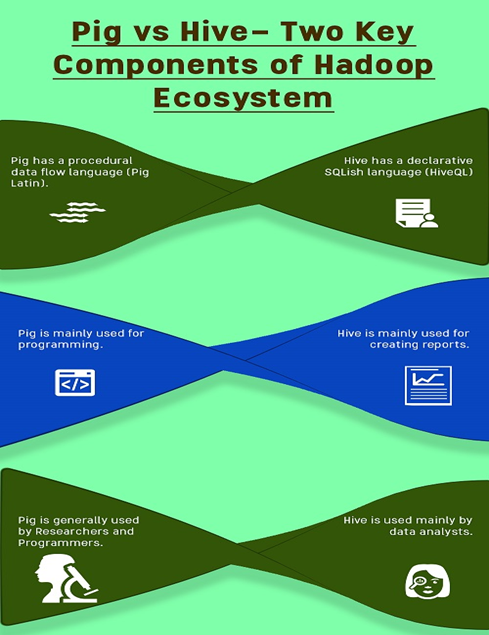
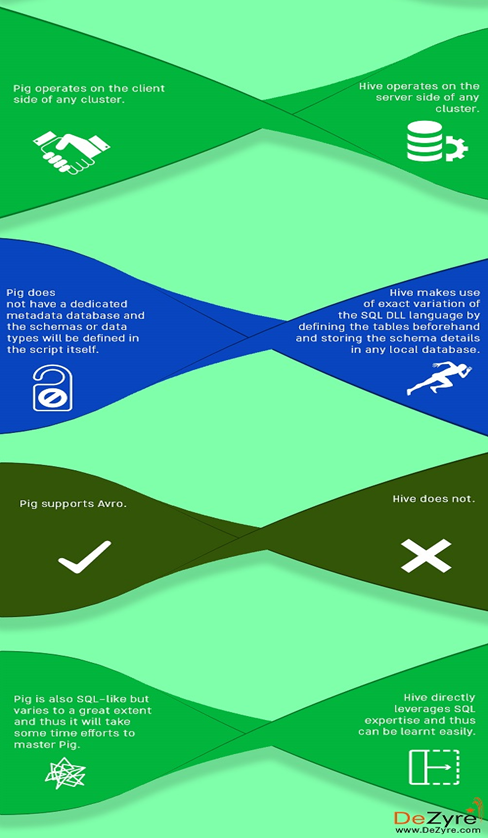
# Difference between Pig and Hive-The Two Key Components of Hadoop Ecosystem

* Pig and Hive are the two key components of the Hadoop ecosystem.
* Hadoop Pig and Hive are similar - they are tools that ease the complexity of writing complex java MapReduce programs.
* Pig vs Hive (Yahoo vs Facebook)

|  |  |
| --- | --- |
| **Pig** | **Hive** |
| Procedural Data Flow Language | Declarative SQL’ish Language |
| For Programming | For creating reports |
| Mainly used by Researchers and Programmers | Mainly used by Data Analysts |
| Operates on the client side of a cluster. | Operates on the server side of a cluster. |
| Does not have a dedicated metadata database. | Makes use of exact variation of dedicated SQL DDL language by defining tables beforehand. |
| Pig is SQL like but varies to a great extent. | Directly leverages SQL and is easy to learn for database experts. |
| Pig supports Avro file format. | Hive does not support AVRO file format. |





Hence Hive Hadoop in points can be summarized as:

* A Data Warehouse Infrastructure
* Definer of a Query Language popularly known as HiveQL (similar to SQL)
* Provides us with various tools for easy extraction, transformation and loading of data.
* Hive allows its users to embed customized mappers and reducers.

Hive Hadoop is very much popular because of following reasons:

* Hive Hadoop provides the users with strong and powerful statistics functions.
* Hive Hadoop is like SQL, so for any SQL developer the learning curve for Hive will almost be negligible.
* Hive Hadoop can be integrated with HBase for querying the data in HBase whereas this is not possible with Pig. In case of Pig, a function named HbaseStorage () will be used for loading the data from HBase.
* Hive Hadoop has gained popularity as it is supported by Hue.
* Hive Hadoop has various user groups such as CNET, Last.fm, Facebook, and Digg and so on.

Pig Hadoop is popular because of following reasons:

* Pig Hadoop follows a multi query approach thus it cuts down on the number times the data is scanned.
* Pig Hadoop is very easy to learn read and write if you are familiar with SQL.
* Pig provides the users with a wide range of nested data types such as Maps, Tuples and Bags that are not present in [MapReduce](http://www.dezyre.com/article/hadoop-mapreduce-vs-apache-spark-who-wins-the-battle/83)along with some major data operations such as Ordering, Filters, and Joins.
* Performance of Pig is on par with the performance of raw Map Reduce.
* Pig has various user groups for instance 90% of Yahoo’s MapReduce is done by Pig, 80% of Twitter’s MapReduce is also done by Pig and various other companies such as Sales force, LinkedIn, AOL and Nokia also employ Pig.

**Differences between Pig and Hive-**

Depending on the purpose and type of data you can either choose to use Hive Hadoop component or Pig Hadoop Component based on the below differences :

1) Hive Hadoop Component is used mainly by data analysts whereas Pig Hadoop Component is generally used by Researchers and Programmers.

2) Hive Hadoop Component is used for completely structured Data whereas Pig Hadoop Component is used for semi structured data.

3) Hive Hadoop Component has a declarative SQLish language (HiveQL) whereas Pig Hadoop Component has a procedural data flow language (Pig Latin)

4) Hive Hadoop Component is mainly used for creating reports whereas Pig Hadoop Component is mainly used for programming.

5) Hive Hadoop Component operates on the server side of any cluster whereas Pig Hadoop Component operates on the client side of any cluster.

6) Hive Hadoop Component is helpful for ETL whereas Pig Hadoop is a great ETL tool for big data because of its powerful transformation and processing capabilities.

7) Hive can start an optional thrift based server that can send queries from any nook and corner directly to the Hive server which will execute them whereas this feature is not available with Pig.

8) Hive directly leverages SQL expertise and thus can be learnt easily whereas Pig is also SQL-like but varies to a great extent and thus it will take some time efforts to master Pig.

9) Hive makes use of exact variation of the SQL DLL language by defining the tables beforehand and storing the schema details in any local database whereas in case of Pig there is no dedicated metadata database and the schemas or data types will be defined in the script itself.

10) The Hive Hadoop component has a provision for partitions so that you can process the subset of data by date or in an alphabetical order whereas Pig Hadoop component does not have any notion for partitions though might be one can achieve this through filters.

11) Pig supports Avro whereas Hive does not.

12) Pig can be installed easily over Hive as it is completely based on shell interaction

13) Pig Hadoop Component renders users with sample data for each scenario and each step through its “Illustrate” function whereas this feature is not incorporated with the Hive Hadoop Component.

14) Hive has smart inbuilt features on accessing raw data but in case of Pig Latin Scripts we are not pretty sure that accessing raw data is as fast as with HiveQL.

15) You can join, order and sort data dynamically in an aggregated manner with Hive and Pig however Pig also provides you an additional COGROUP feature for performing outer joins.