

Virtuoso Infotech Pvt. Ltd.



About Virtuoso Infotech

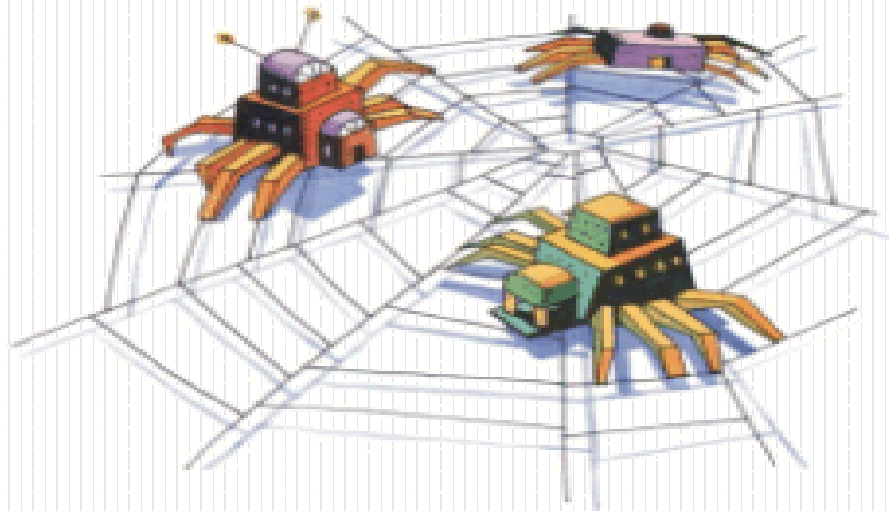
- Fastest growing IT firm; Offers the flexibility of a small firm and robustness of over 30 years experience collectively within the leadership team
- Technology expertise & passionate team
- Successful client engagements across India, USA, UK, Australia and Argentina
- Handle enterprise solutions that involve **30,000 active users**, more than 20 servers, **data volume as big as 5 million entries per day**

Agenda

- An Introduction of Hadoop.
- Sub Projects of Hadoop.
- An Architecture of Hadoop.
- How Does Hadoop Work?
- Advantages of Hadoop.
- Disadvantages of Hadoop.

An Introduction of Hadoop

Designed to answer the question: **“How to process big data with reasonable cost and time?”**



An Introduction of Hadoop

- Open-source software framework for distributed storage and distributed processing of very large data sets on computer clusters built from commodity hardware.
- Created by Doug Cutting (Creator of Apache Lucene, widely used text search library.)
- Has its origins in Apache Nutch, an open source web search engine.
- Uses Google's MapReduce Algorithm.

Sub Projects of Hadoop

Pig	Chukwa	Hive	HBase
MapReduce	HDFS	ZooKeeper	
Core		Avro	

An Architecture of Hadoop

Map Reduce
(Distributed Computation)

HDFS
(Distributed Storage)

Yarn
Framework

Common
Utility

How Does Hadoop Work?

This process includes the following core tasks that Hadoop performs:

- Data is initially divided into directories and files. Files are divided into uniform sized
- blocks of 128M and 64M (preferably 128M). These files are then distributed across various cluster nodes for further processing.
- HDFS, being on top of the local file system, supervises the processing.

How Does Hadoop Work?

- Blocks are replicated for handling hardware failure.
- Checking that the code was executed successfully.
- Performing the sort that takes place between the map and reduce stages.
- Sending the sorted data to a certain computer.
- Writing the debugging logs for each job.

Advantages of Hadoop?

- Distributed Data.
- Independent Task.
- Linear scaling.
- Simple programming model.
- Flat scalability.
- HDFS store large simple and robust coherency model.
- Can be offered as an on-demand service.

Disadvantages of Hadoop

- Rough manner.
- Programming model is very restrictive.
- Joins of multiple datasets are tricky and slow.
- Cluster management is hard.
- Still single master which requires care and may limit scaling
- Managing job flow isn't trivial when intermediate data should be kept
- Optimal configuration.

Thank You!

Virtuoso InfoTech Pvt. Ltd.
4th Floor, Victory Landmark, Opp. D-
Mart,
Behind Dominos Pizza, Baner, Pune.

+91 8087081318
support@virtuosoitech.com



www.virtuosoitech.com

