This book jumps into the world of Hadoop ecosystem components and its tools in a simplified manner, and provides you with the skills to utilize them effectively for faster and effective development of Hadoop projects.

Starting with the concepts of Hadoop YARN, MapReduce, HDFS, and other Hadoop ecosystem components, you will soon learn many exciting topics such as MapReduce patterns, data management, and real-time data analysis using Hadoop. You will also get acquainted with many Hadoop ecosystem components tools such as Hive, HBase, Pig, Sqoop, Flume, Storm, and Spark.

By the end of the book, you will be confident to begin working with Hadoop straightforward and implement the knowledge gained in all your real-world scenarios.

Who this book is written for

If you are a system or application developer interested in learning how to solve practical problems using the Hadoop framework, then this book is ideal for you. This book is also meant for Hadoop professionals who want to find solutions to the different challenges they come across in their Hadoop projects.

What you will learn from this book

- Get introduced to Hadoop, big data, and the pillars of Hadoop such as HDFS, MapReduce, and YARN
- Understand different use cases of Hadoop along with big data analytics and real-time analysis in Hadoop
- Explore the Hadoop ecosystem tools and effectively use them for faster development and maintenance of a Hadoop project
- Demonstrate YARN's capacity for database processing
- Work with Hive, HBase, and Pig with Hadoop to easily figure out your big data problems
- Gain insights into widely used tools such as Sqoop, Flume, Storm, and Spark using practical examples

Delve into the key concepts of Hadoop and get a thorough understanding of the Hadoop ecosystem.
In this package, you will find:

- The author biography
- A preview chapter from the book, Chapter 2 'Hadoop Ecosystem'
- A synopsis of the book’s content
- More information on Hadoop Essentials

About the Author

Shiva Achari has over 8 years of extensive industry experience and is currently working as a Big Data Architect consultant with companies such as Oracle and Teradata. Over the years, he has architected, designed, and developed multiple innovative and high-performance large-scale solutions, such as distributed systems, data centers, big data management tools, SaaS cloud applications, Internet applications, and Data Analytics solutions.

He is also experienced in designing big data and analytics applications, such as ingestion, cleansing, transformation, correlation of different sources, data mining, and user experience in Hadoop, Cassandra, Solr, Storm, R, and Tableau.

He specializes in developing solutions for the big data domain and possesses sound hands-on experience on projects migrating to the Hadoop world, new developments, product consulting, and POC. He also has hands-on expertise in technologies such as Hadoop, Yarn, Sqoop, Hive, Pig, Flume, Solr, Lucene, Elasticsearch, Zookeeper, Storm, Redis, Cassandra, HBase, MongoDB, Talend, R, Mahout, Tableau, Java, and J2EE.

He has been involved in reviewing Mastering Hadoop, Packt Publishing. Shiva has expertise in requirement analysis, estimations, technology evaluation, and system architecture along with domain experience in telecoms, Internet applications, document management, healthcare, and media.

Currently, he is supporting presales activities such as writing technical proposals (RFP), providing technical consultation to customers, and managing deliveries of big data practice groups in Teradata.

He is active on his LinkedIn page at http://in.linkedin.com/in/shivaachari/.
Hadoop Essentials

Hadoop is quite a fascinating and interesting project that has seen quite a lot of interest and contributions from the various organizations and institutions. Hadoop has come a long way, from being a batch processing system to a data lake and high-volume streaming analysis in low latency with the help of various Hadoop ecosystem components, specifically YARN. This progress has been substantial and has made Hadoop a powerful system, which can be designed as a storage, transformation, batch processing, analytics, or streaming and real-time processing system.

Hadoop project as a data lake can be divided in multiple phases such as data ingestion, data storage, data access, data processing, and data management. For each phase, we have different sub-projects that are tools, utilities, or frameworks to help and accelerate the process. The Hadoop ecosystem components are tested, configurable and proven and to build similar utility on our own it would take a huge amount of time and effort to achieve. The core of the Hadoop framework is complex for development and optimization. The smart way to speed up and ease the process is to utilize different Hadoop ecosystem components that are very useful, so that we can concentrate more on the application flow design and integration with other systems.

With the emergence of many useful sub-projects in Hadoop and other tools within the Hadoop ecosystem, the question that arises is which tool to use when and how effectively. This book is intended to complete the jigsaw puzzle of when and how to use the various ecosystem components, and to make you well aware of the Hadoop ecosystem utilities and the cases and scenarios where they should be used.
What This Book Covers

Chapter 1, Introduction to Big Data and Hadoop, covers an overview of big data and Hadoop, plus different use case patterns with advantages and features of Hadoop.

Chapter 2, Hadoop Ecosystem, explores the different phases or layers of Hadoop project development and some components that can be used in each layer.

Chapter 3, Pillars of Hadoop – HDFS, MapReduce, and YARN, is about the three key basic components of Hadoop, which are HDFS, MapReduce, and YARN.

Chapter 4, Data Access Components – Hive and Pig, covers the data access components Hive and Pig, which are abstract layers of the SQL-like and Pig Latin procedural languages, respectively, on top of the MapReduce framework.

Chapter 5, Storage Components – HBase, is about the NoSQL component database HBase in detail.

Chapter 6, Data Ingestion in Hadoop – Sqoop and Flume, covers the data ingestion library tools Sqoop and Flume.

Chapter 7, Streaming and Real-time Analysis – Storm and Spark, is about the streaming and real-time frameworks Storm and Spark built on top of YARN.
Now that we have discussed and understood big data and Hadoop, we can move on to understanding the Hadoop ecosystem. A Hadoop cluster may have hundreds or thousands of nodes which are difficult to design, configure, and manage manually. Due to this, there arises a need for tools and utilities to manage systems and data easily and effectively. Along with Hadoop, we have separate sub-projects which are contributed by some organizations and contributors, and are managed mostly by Apache. The sub-projects integrate very well with Hadoop and can help us concentrate more on design and development rather than maintenance and monitoring, and can also help in the development and data management.

Before we understand different tools and technologies, let's understand a use case and how it differs from traditional systems.

**Traditional systems**

Traditional systems are good for OLTP (online transaction processing) and some basic Data Analysis and BI use cases. Within the scope, the traditional systems are best in performance and management. The following figure shows a traditional system on a high-level overview:
The steps for typical traditional systems are as follows:

1. Data resides in a database
2. ETL (Extract Transform Load) processes
3. Data moved into a data warehouse
4. Business Intelligence Applications can have some BI reporting
5. Data can be used by Data Analysis Application as well

When the data grows, traditional systems fail to process, or even store, the data; and even if they do, it comes at a very high cost and effort because of the limitations in the architecture, issue with scalability and resource constraints, incapability or difficulty to scale horizontally.

**Database trend**

Database technologies have evolved over a period of time. We have RDBMS (relational database), EDW (Enterprise data warehouse), and now Hadoop and NoSQL-based database have emerged. Hadoop and NoSQL-based database are now the preferred technology used for the big data problems, and some traditional systems are gradually moving towards Hadoop and NoSQL, along with their existing systems. Some systems have different technologies to process the data such as, Hadoop with RDBMS, Hadoop with EDW, NoSQL with EDW, and NoSQL with Hadoop. The following figure depicts the database trend according to Forrester Research:

The figure depicts the design trends and the technology which was available and adapted in a particular decade.

The 1990's decade was the RDBMS era which was designed for OLTP processing and data processing was not so complex.
The emergence and adaptation of data warehouse was in the 2000's, which is used for OLAP processing and BI.

From 2010 big data systems, especially Hadoop, have been adapted by many organizations to solve Big Data problems.

All these technologies can practically co-exist for a solution as each technology has its pros and cons because not all problems can be solved by any one technology.

The Hadoop use cases

Hadoop can help in solving the big data problems that we discussed in Chapter 1, Introduction to Big Data and Hadoop. Based on Data Velocity (Batch and Real time) and Data Variety (Structured, Semi-structured and Unstructured), we have different sets of use cases across different domains and industries. All these use cases are big data use cases and Hadoop can effectively help in solving them. Some use cases are depicted in the following figure:
Hadoop's basic data flow

A basic data flow of the Hadoop system can be divided into four phases:

1. **Capture Big Data**: The sources can be extensive lists that are structured, semi-structured, and unstructured, some streaming, real-time data sources, sensors, devices, machine-captured data, and many other sources. For data capturing and storage, we have different data integrators such as, Flume, Sqoop, Storm, and so on in the Hadoop ecosystem, depending on the type of data.

2. **Process and Structure**: We will be cleansing, filtering, and transforming the data by using a MapReduce-based framework or some other frameworks which can perform distributed programming in the Hadoop ecosystem. The frameworks available currently are MapReduce, Hive, Pig, Spark and so on.

3. **Distribute Results**: The processed data can be used by the BI and analytics system or the big data analytics system for performing analysis or visualization.

4. **Feedback and Retain**: The data analyzed can be fed back to Hadoop and used for improvements and audits.

The following figure shows the data captured and then processed in a Hadoop platform, and the results used in a Business Transactions and Interactions system, and a Business Intelligence and Analytics system:
**Hadoop integration**

Hadoop architecture is designed to be easily integrated with other systems. Integration is very important because although we can process the data efficiently in Hadoop, but we should also be able to send that result to another system to move the data to another level. Data has to be integrated with other systems to achieve interoperability and flexibility.

The following figure depicts the Hadoop system integrated with different systems and with some implemented tools for reference:

![Hadoop Integration with other systems](image)

Systems that are usually integrated with Hadoop are:

- Data Integration tools such as, Sqoop, Flume, and others
- NoSQL tools such as, Cassandra, MongoDB, Couchbase, and others
- ETL tools such as, Pentaho, Informatica, Talend, and others
- Visualization tools such as, Tableau, Sas, R, and others

**The Hadoop ecosystem**

The Hadoop ecosystem comprises of a lot of sub-projects and we can configure these projects as we need in a Hadoop cluster. As Hadoop is an open source software and has become popular, we see a lot of contributions and improvements supporting Hadoop by different organizations. All the utilities are absolutely useful and help in managing the Hadoop system efficiently. For simplicity, we will understand different tools by categorizing them.
Hadoop Ecosystem

The following figure depicts the layer, and the tools and utilities within that layer, in the Hadoop ecosystem:

Distributed filesystem

In Hadoop, we know that data is stored in a distributed computing environment, so the files are scattered across the cluster. We should have an efficient filesystem to manage the files in Hadoop. The filesystem used in Hadoop is HDFS, elaborated as Hadoop Distributed File System.

HDFS

HDFS is extremely scalable and fault tolerant. It is designed to efficiently process parallel processing in a distributed environment in even commodity hardware. HDFS has daemon processes in Hadoop, which manage the data. The processes are NameNode, DataNode, BackupNode, and Checkpoint NameNode.

We will discuss HDFS elaborately in the next chapter.
Where to buy this book

You can buy Hadoop Essentials from the Packt Publishing website.

Alternatively, you can buy the book from Amazon, BN.com, Computer Manuals and most internet book retailers.

Click here for ordering and shipping details.

Get more information Hadoop Essentials