**Project: Apache Hadoop – Airline Dataset analysis**

**TECHNOLOGY - HADOOP**

The fundamentals of MapReduce are described here by using practice examples. RDBMS and SQL are ubiquitous in data-processing environments; we use language elements from SQL to explain the basic concepts of MapReduce. To achieve the goals of this chapter, we use an airline dataset that comprises arrival and departure flight details of all commercial flights from 1987 through 2008. We introduce this dataset first; you can then start tackling familiar data-processing problems that we mentioned in terms of SQL language elements by using MapReduce. In the process, you will gain familiarity with various components of MapReduce, including these:

* Mapper
* Reducer
* Combiner
* Partitioner

**Reviewing the Airline Dataset**

This project uses an airline dataset that consists of flight arrival and departure details of commercial domestic flights in the United States from 1987 to 2008. This dataset holds approximately 120 million records in comma-separated value (CSV) format. The total disk size of this dataset in uncompressed format is 120 GB. Although it is a large dataset that is suitable for Hadoop processing, it is not too large, so it is suitable for the purpose of this book. The other reason for selecting this dataset is to discuss MapReduce in the context of structured data processing. Although it is true that Hadoop is used for unstructured data processing, its most common use is to perform large-scale ETL on structured datasets and use it as a data warehouse for large datasets. The airline dataset is highly structured and suitable for explaining MapReduce concepts in terms of SQL language elements.

This project explored all the major components of MapReduce using a use-case driven approach. We took the familiar SQL paradigm and developed common data processing functions such as SELECT, WHERE, and GROUP BY aggregations; so the basics of MapReduce development is learned as it applies to real-world situations and large-scale datasets.