MODULE III PART 3
MapReduce

At the end of the session you will be able

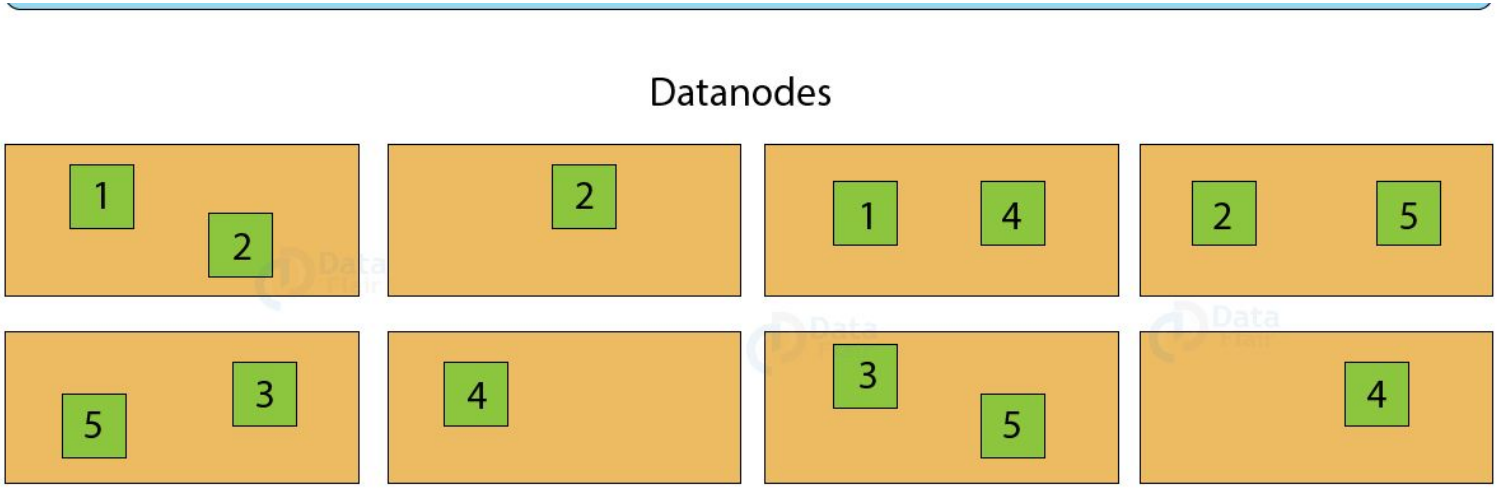
- Explain MapReduce.

HDFS

Replication Management

- To provide fault tolerance
- Copies of data blocks are stored in different DataNodes.
- 3 by default.Can be configured to any value.
- If there is a file with size 1GB then with a replication factor 3, it will require 3GB storage

HDFS -Replication

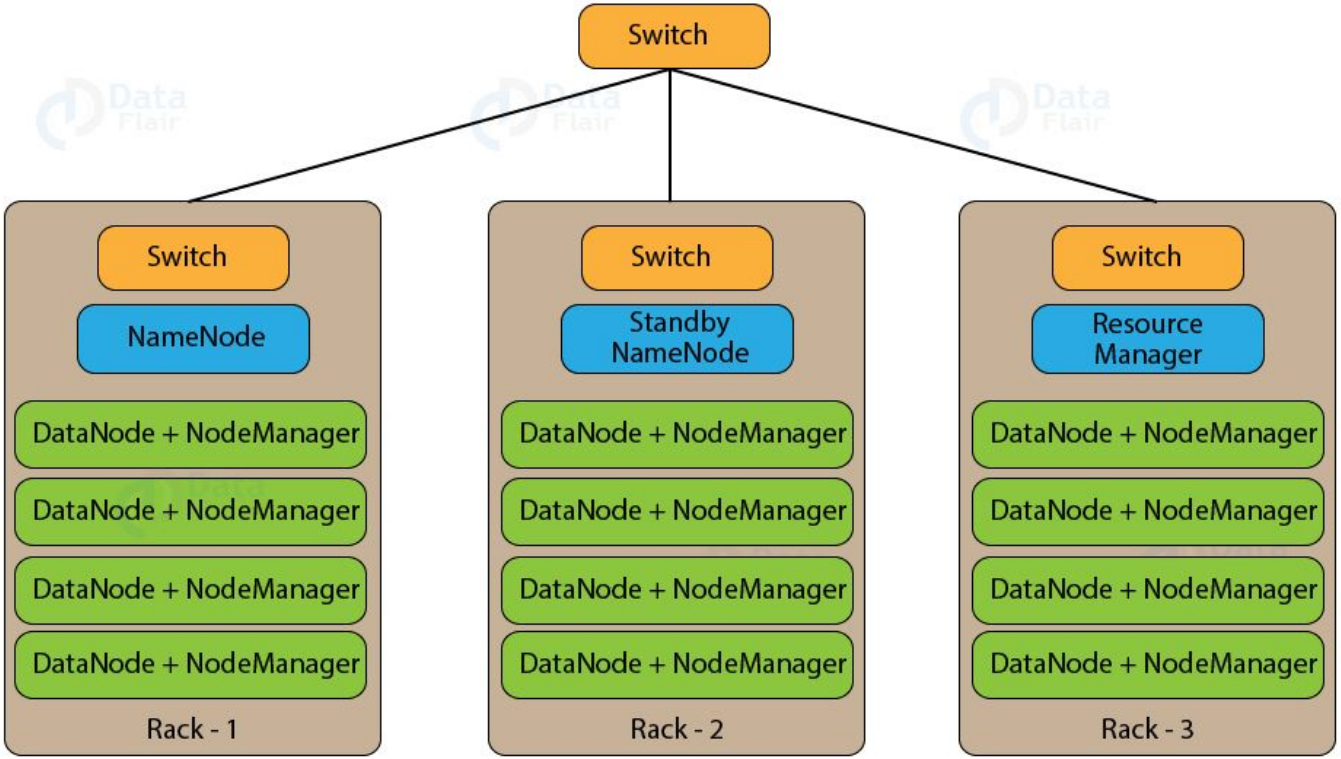


HDFS

Rack Awareness

- A rack contains several nodes.
- There are several racks.
- Rack awareness algorithm for distributing blocks.
- Provides low latency and fault tolerance.

Rack Awareness



- **Data Processing Layer.**
- **Works by breaking the processing into two phases:**
  - **Map phase**
  - **Reduce phase**
- **Each phase has key-value pairs.**
- **Programmer also defines two functions:**
  - **Map function**
  - **Reduce function**

- **Map function**
  - Load, parse, transform and filter data.
  - Reduce task works on the sub-set of output from the map tasks.
- **Reduce function**
  - Applies grouping and aggregation to intermediate data from the map tasks.

EXAMPLE-WEATHER DATASET

- Aim
  - To find the highest recorded global temperature for each year.
- Characteristics of Data
  - Huge dataset.
  - Stored in line-oriented ASCII format.
  - Each line is a record.
  - Source:National Climatic Data Center (NCDC, http://www.ncdc.noaa.gov/)

EXAMPLE-WEATHER DATASET

- Map phase
  - Input-raw NCDC data
- Map function
  - Data preparation Phase.
  - Pull out the year and the air temperature.
  - Drops bad records; filter out temperatures that are missing

EXAMPLE-WEATHER DATASET

Sample data

```
0067011990999991950051507004...9999999N9+00001+99999999999...
0043011990999991950051512004...9999999N9+00221+99999999999...
0043011990999991950051518004...9999999N9-00111+99999999999...
0043012650999991949032412004...0500001N9+01111+99999999999...
0043012650999991949032418004...0500001N9+00781+99999999999...
```

(some unused columns have been dropped to fit the page, indicated by ellipses)

These lines are presented to the map function as the key-value pairs:

```
(0, 0067011990999991950051507004...9999999N9+00001+99999999999...)
(106, 0043011990999991950051512004...9999999N9+00221+99999999999...)
(212, 0043011990999991950051518004...9999999N9-00111+99999999999...)
(318, 0043012650999991949032412004...0500001N9+01111+99999999999...)
(424, 0043012650999991949032418004...0500001N9+00781+99999999999...)
```

EXAMPLE-WEATHER DATASET

- The map function merely extracts the year and the air temperature.

```
(1950, 0)
(1950, 22)
(1950, -11)
(1949, 111)
(1949, 78)
```

- The output from the map function is processed by the MapReduce framework.
- This processing sorts and groups the key-value pairs by key.

```
(1949, [111, 78])
(1950, [0, 22, -11])
```

EXAMPLE-WEATHER DATASET

- Reduce Function
  - Iterate through the list and pick up the maximum reading
  -

```
(1949, 111)
(1950, 22)
```