AJAY KUMAR GARG ENGINEERING COLLEGE, GHAZIABAD

Computer Science And Engineering

Data Analytics and Visualization (KDS-501)

Data Analytics LifeCycle

Presented By:
Ms. Neeharika Tripathi
Assistant Professor
Department of Computer Science And Engineering

Cloud Computing

- Cloud computing is on-demand access, via the internet, to computing resources—applications, servers (physical servers and virtual servers), data storage, development tools, networking capabilities, and more—hosted at a remote data center managed by a cloud services provider (or CSP). The CSP makes these resources available for a monthly subscription fee or bills them according to usage.
- Public and Private Cloud: In a private cloud, a single organization controls and maintains the underlying infrastructure to deliver the IT resources. In a public cloud, external cloud providers deliver the resources as a fully

Grid computing

- Grid computing is a computing infrastructure that combines computer resources spread over different geographical locations to achieve a common goal. All unused resources on multiple computers are pooled together and ma k.

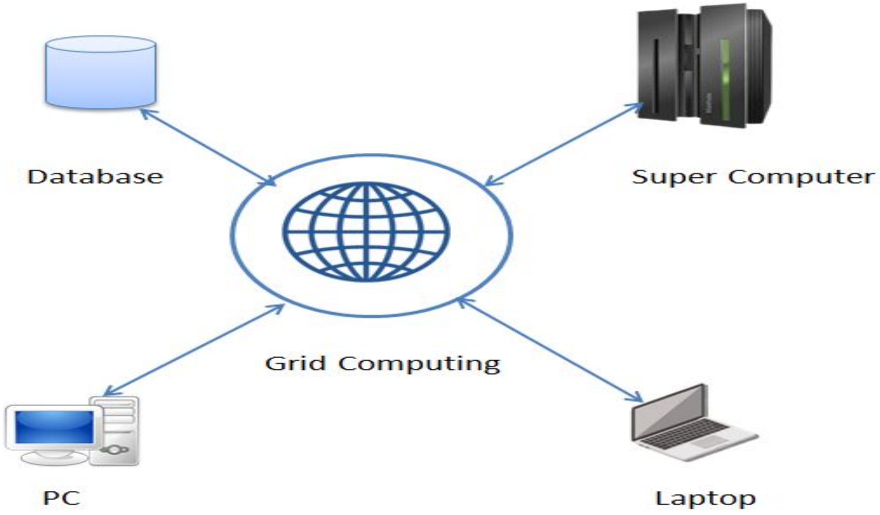


Fig: Grid Computing Architecture

Components of Grid Computing:

- Nodes
- The computers or servers on a grid computing network are called nodes. Each node offers unused computing resources such as CPU, memory, and storage to the grid network.
- Grid middleware

Grid middleware is a specialized software application that connects computing resources in grid operations with high-level applications. It controls the user sharing of available resources to prevent overwhelming the grid computers. The grid middleware also provides security to prevent misuse of resources in grid

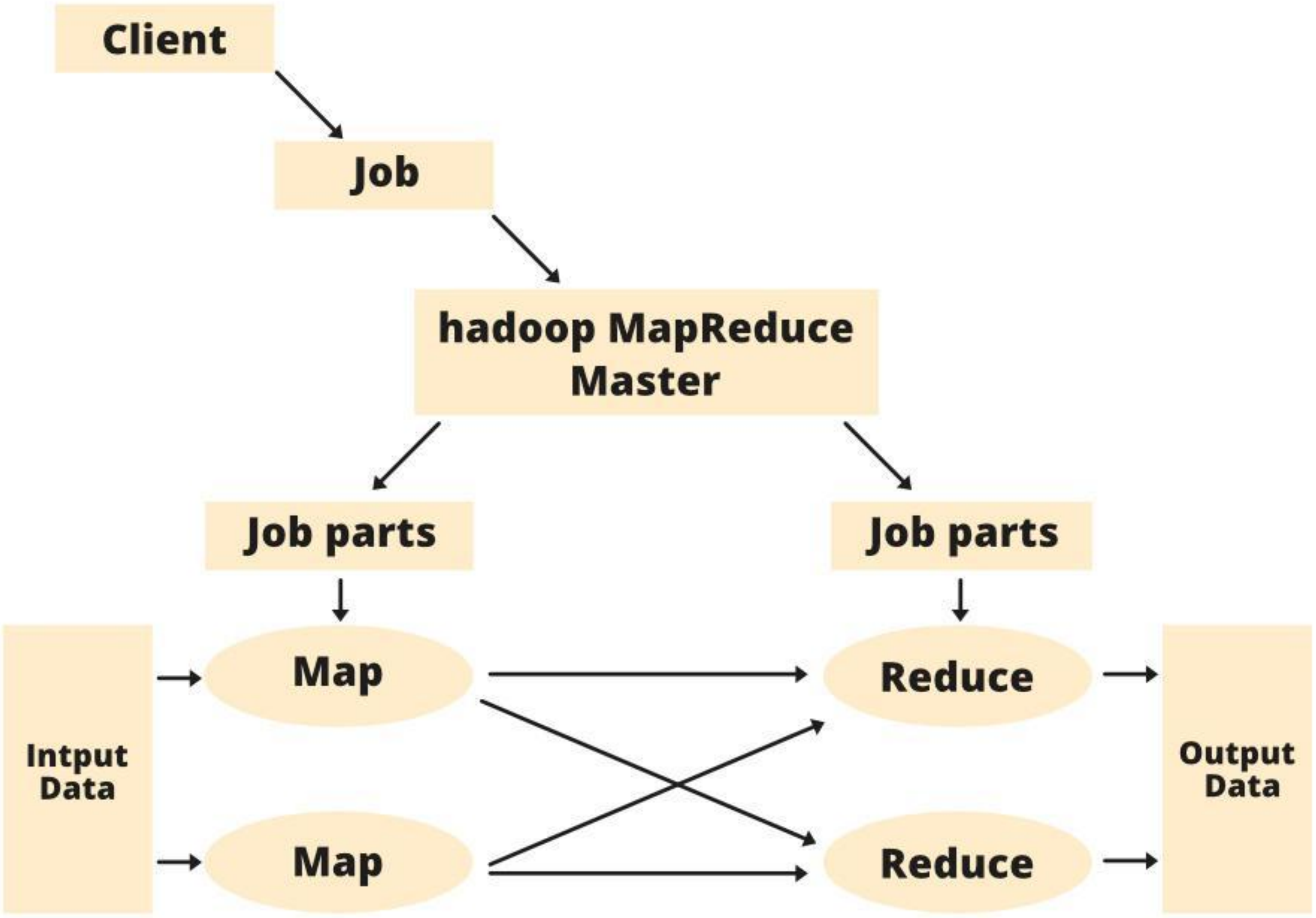
Grid computing architecture

- Grid architecture represents the internal structure of grid computers. The following layers are broadly present in a grid node:
- The top layer consists of high-level applications, such as an application to perform predictive modeling.
- The second layer, also known as middleware, manages and allocates resources requested by applications.
- The third layer consists of available computer resources such as CPU, memory, and storage.
- The bottom layer allows the computer to connect to a grid computing network.

Map Reduce

- MapReduce is a programming model used for efficient processing in parallel over large data-sets in a distributed manner. The data is first split and then combined to produce the final result.
- The purpose of MapReduce in Hadoop is to Map each of the jobs and then it will reduce it to equivalent tasks for providing less overhead over the cluster network and to reduce the processing power.

Map Reduce Architecture



Components of MapReduce architecture

- **Client:** The MapReduce client is the one who brings the Job to the MapReduce for processing.
- **Job:** The MapReduce Job is the actual work that the client wanted to do which is comprised of so many smaller tasks
- **Hadoop MapReduce Master:** It divides the particular job into subsequent job-parts.
- **Job-Parts:** The task or sub-jobs that are obtained after dividing the main job. The result of all the job-parts combined to produce the final output.
- **Input Data:** The data set that is fed to the MapReduce for processing.
- **Output Data:** The final result is obtained after

Key Roles in Data Analytics Projects

- There are certain key roles that are required for the complete and fulfilled functioning of the data science team to execute projects on analytics successfully. The key roles are seven in number.
- Each key plays a crucial role in developing a successful analytics project.

1. **Business User** :The business user is the one who understands the main area of the project and is also basically benefited from the results.

- This user gives advice and consult the team working on the project about the value of the results obtained and how the operations on the outputs are done.
- The business manager, line manager, or deep

2. **Project Sponsor :**

- The Project Sponsor is the one who is responsible to initiate the project. Project Sponsor provides the actual requirements for the project and presents the basic business issue.
- He generally provides the funds and measures the degree of value from the final output of the team working on the project.
- This person introduce the prime concern and brooms the desired output.

3. **Project Manager :**This person ensures that key milestone and purpose of the project is met on time and of the expected quality.

4. **Business Intelligence Analyst :**Business Intelligence Analyst provides business domain perfection based on a detailed and deep understanding of the data, key performance indicators (KPIs), key matrix, and business intelligence from a reporting point of view.

5. Database Administrator (DBA) :

- DBA facilitates and arrange the database environment to support the analytics need of the team working on a project.
- His responsibilities may include providing permission to key databases or tables and making sure that the appropriate security stages are in their correct places related to the data repositories or not.

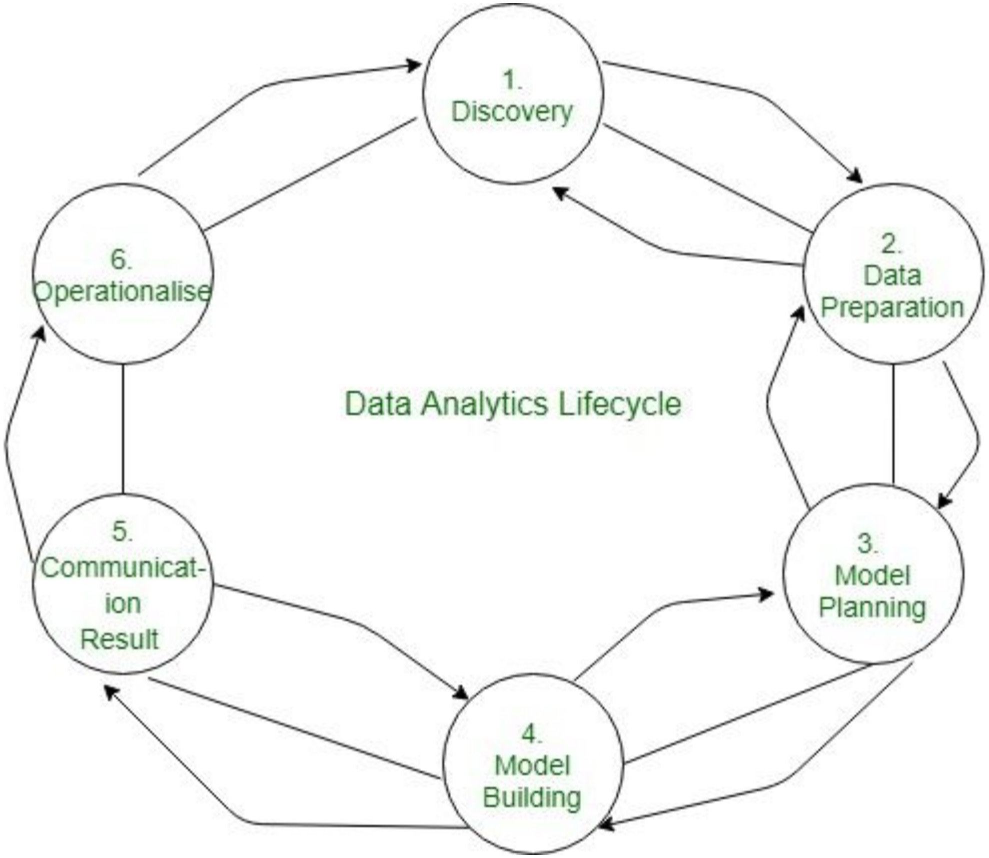
6. Data Engineer :

- Data engineer grasps deep technical skills to assist with tuning SQL queries for data management and data extraction and provides support for data intake into the analytic sandbox.
- The data engineer works jointly with the data scientist to help build data in correct ways for analysis.

7. Data Scientist :

- Data scientist facilitates with the subject matter expertise for analytical techniques, data modelling, and applying correct analytical techniques for a given business issues.
- He ensures overall analytical objectives are met.

Various Phases in Data Analytics Life Cycle



1. **Phase 1: Discovery –**The data science team learn and investigate the problem.

- Develop context and understanding.
- Come to know about data sources needed and available for the project.
- The team formulates initial hypothesis that can be later tested with data.

2. **Phase 2: Data Preparation-** Steps to explore, preprocess, and condition data prior to modeling and analysis.

- It requires the presence of an analytic sandbox, the team execute, load, and transform, to get data into the sandbox.
- Data preparation tasks are likely to be performed multiple times and not in predefined order.

- **3. Phase 3: Model Planning** –Team explores data to learn about relationships between variables and subsequently, selects key variables and the most suitable models.
- In this phase, data science team develop data sets for training, testing, and production purposes.
- Team builds and executes models based on the work done in the model planning phase.
- Several tools commonly used for this phase are – Matlab, STASTICA.

4. Phase 4: Model Building –Team develops datasets for testing, training, and production purposes.

- Team also considers whether its existing tools will suffice for running the models or if they need more robust environment for executing models.
- Free or open-source tools – Rand PL/R, Octave, WEKA

5. **Phase 5: Communication Results –**After executing model team need to compare outcomes of modeling to criteria established for success and failure.

- Team considers how best to articulate findings and outcomes to various team members and stakeholders, taking into account warning, assumptions.
- Team should identify key findings, quantify business value, and develop narrative to summarize and convey findings to stakeholders.

6. **Phase 6: Operationalize –**The team communicates benefits of project more broadly and sets up pilot project to deploy work in controlled way before broadening the work to full enterprise of users.

- This approach enables team to learn about performance and related constraints of the model in production environment on small scale , and make adjustments before full deployment.
- The team delivers final reports, briefings, codes