https://github.com/JyotinderSingh

PROGRAMMING MODELS FOR BIG DATA:

1. Support Big Data Operations
   - Split volumes of data
   - Access data fast
   - Distribute computations to nodes
2. Handle fault Tolerance
   - Replicate Data Partitions
   - Recover files when needed
3. Enable adding more Racks/Resources
4. Optimized for specific data types
   - Document, Table, key-value, Graphs, Multimedia, stream.

MAP REDUCE

↓

A PROGRAMMING MODEL FOR BIG DATA

↓

MANY IMPLEMENTATIONS

https://github.com/JyotinderSingh

THE HADOOP ECOSYSTEM

HDFS : Hadoop Distributed File System → Scalable Storage, fault tolerance
↳ Yahoo

Yarn : Flexible Scheduling and resource management

MapReduce : Programming model that simplifies parallel computing
↳ Google

Map → apply()

Reduce → summarize()

Pig : Dataflow scripting
↳ Yahoo

Hive : SQL-like queries
↳ facebook

Giraph : built for processing large scale graphs efficiently
↳ facebook uses it to analyse the social graphs of its users.

Storm
Spark
Flink
] Real time & in-memory processing

HBase → used in facebook's messaging platform
Cassandra
MongoDB
] - NoSQL for non-files
- key-values
- Sparse tables

Cassandra ↳ created at facebook

https://github.com/JyotinderSingh

Zookeeper : for management
- synchronization
- configeration
- high availability

created by Yahoo to wrangle services named after animals