Hadoop Core

Tuesday, January 10, 2023 7:09 AM

- Requirements
  - Data Capture
    - Handle ingestion of both structured/unstructured data from multiple data sources (quickly/efficiently)
    - Handle both batch & online use cases

      warehouse/reporting realtime/streaming
  - Data Processing
    - ETL: Extract/Transform/Load
    - OLAP: Aggregation, Queries, Reports (Analytics)

      Batch

      Ad-Hoc (Interactive)
  - Data Exchange
    - Integration: Interoperability with other systems
    - Sharing: Data/Results across systems & students
  - Easy to Operate
    - Administration: Provisioning, Scaling, Diagnosing, Monitoring, Managing, Maintaining
    - Operations: Availability, Reliability, Dirt Testing....

* Conceptually

Temporal

Scale {
- Compute: Code (Task) is "too large" - run distributed across a system
- Storage: Data (Memory) is "too large" - store distributed across a system

Spatial

- Categories
  - MPP (Massively Parallel Processing)
    - Hadoop ✓
    - ELK: Elasticsearch, Logstash, Kibana (Lucene/Solr) => Document Search/IR
    - Postgres + PostGIS & Friends => Geospatial
    - Clickhouse & Friends -> Time Series

    * Row stores (OLTP)
  - NoSQL (Non-Relational)
    - CouchDB, MongoDB => Document DB
    - Cassandra, HBase => Wide-Column DB
    - Memcached, Redis => Key-Value DB
    - BigTable, DynamoDB, CosmosDB -> Multi Paradigm DB
    - Neo4j, Gremlin => Graph DB
  - Column Stores (OLAP)
    - CStore => Vertica
    - Redshift

- Monet DB
- kdb+

* HPC Systems

OpenMPI & Friends → Supercomputing / Scientific computing

- PFS, OpenMP, OpenACC ....

- Hadoop
  - Concepts (Files/Blocks)
    - Storage: Data loaded into system → Distribute across nodes
    - Compute: Code (Jobs/Tasks) run on system → Distribute across nodes

↑ Where the data is located!

* Locality

- Techniques
  - COTS Hardware: Linux / x86-64 / Ethernet / Disk
    - CPU, ENet → HDD (PCIE)
  - API: Understandable / Usable by most programmers

Open Source (Dev)

↕

Open Access (User)

* Hadoop Core & Hadoop Ecosystem are two different things!

- Characteristics
  - Scalable
  - Reliability
  - Flexibility
  - Economical

- History
  - Google
    - Crawl / Index documents
    - PageRank Calculations → Linear Algebra / Matrix-Vector
  - Proprietary
    - GFS
    - MR

* Open-Source Implementation
- Apache
- Doug Cutting

- Uptake
  - · IT / Cost
    - Storage Costs
    - ETL Efficiency (TTM)
    - EDW Optimization
    - Archiving
  - · Business / Revenue
    - Data Science
    - EDA "Business Analytics"
    - Predictive Analytics / Data Mining

- Hadoop Core
  - · Architecture
    - Common : Libraries, Utilities, etc...
    - storage - HDFS : FOSS implementation of GFS ⇒ DFS
      ↳ runs across all Hadoop cluster nodes!
    - compute - YARN : Cluster resource management & monitoring
      ↳ utilization / availability ⇒ intelligent / efficient job (task) scheduling
    - Zookeeper : Centralized service holder cluster state data
      ↳ DHT serves as KV-store