Big Data Analysis

Lecture 4

2019/11/25

Data Storage Choices

- Relational databases
- NoSQL databases
- Time Series Databases

Relational Databases

- Review of SQL and Queries
- Schema Examples from Wind WDS Portal
- Show performance comparison between SQL from sqlPlus and Matlab programs
- Will creating views help?

Discussions

- Domain specific applications require more convenient data structures for reasonable performance.
- Domain specific knowledge needs to be taken into account.
- Matlab
  - Structures (struct) are very versatile, mixture of key-value pairs and hierarchical structures.
  - Flexibility to add new field and new column, accommodate higher dimension data …
  - Financial time series (fints)
  - Map container (map)

NoSQL Databases

- NoSQL – Not Only SQL
- Handle big data – 3 V's
- No predefined schemas
- Cheaper to manage storage, but …
- Easy and cheap to scale horizontally

Advantage of Relational Databases

- Better for relational data and models
- Normalization – efficiency
- SQL – simple, solid, efficient
- Data integrity
- ACID compliance

NoSQL Database Examples

- Document databases: MongoDB
- Key-Value databases: Redis
- Graph databases
- Time Series databases: InfluxDB, KDB+

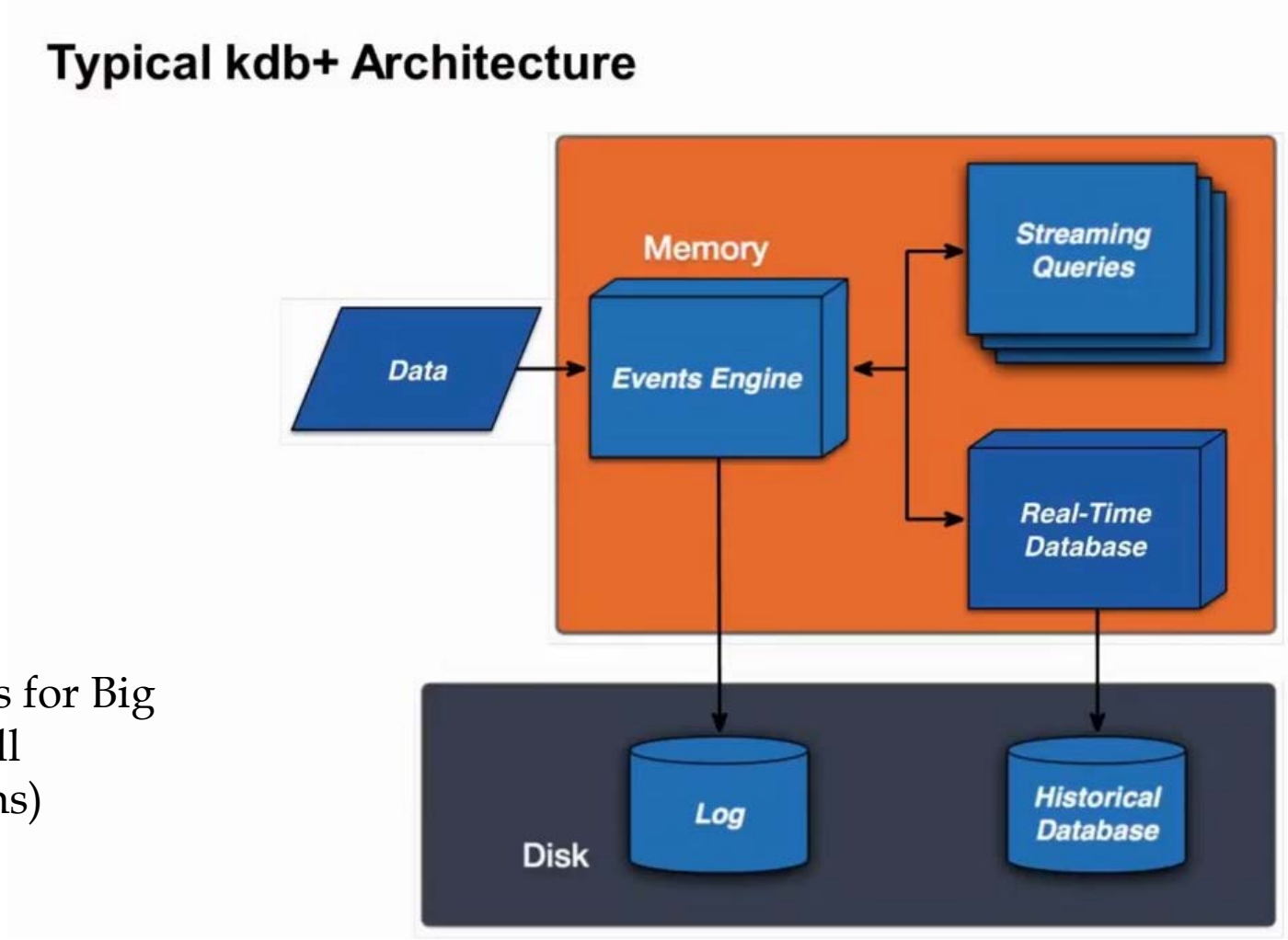
MongoDB

- https://github.com/mongodb/mongo
- Document
  - https://docs.mongodb.com/manual/
- More reading materials
  - https://university.mongodb.com/
- Installation guide
  - https://www.cnblogs.com/TM0831/p/10606624.html
- Demos
  - https://gist.github.com/bradtraversy/f407d642bdc3b31681bc7e56d95485b6

Time Series Databases

- Introduction from InfluxDB
  - https://www.influxdata.com/time-series-database/
- Time is treated as first class with finest granularity
- Time-window based operations are highly optimized
- Column vector operation
- Mostly read and append operations

KDB+ Architecture



Time Series Analytics for Big Fast Data Fintan Quill (Engineer, Kx Systems) October 26, 2017