Time Series Database

Team 9

Balamanikandan Gopalakrishnan

Rachan Hegde

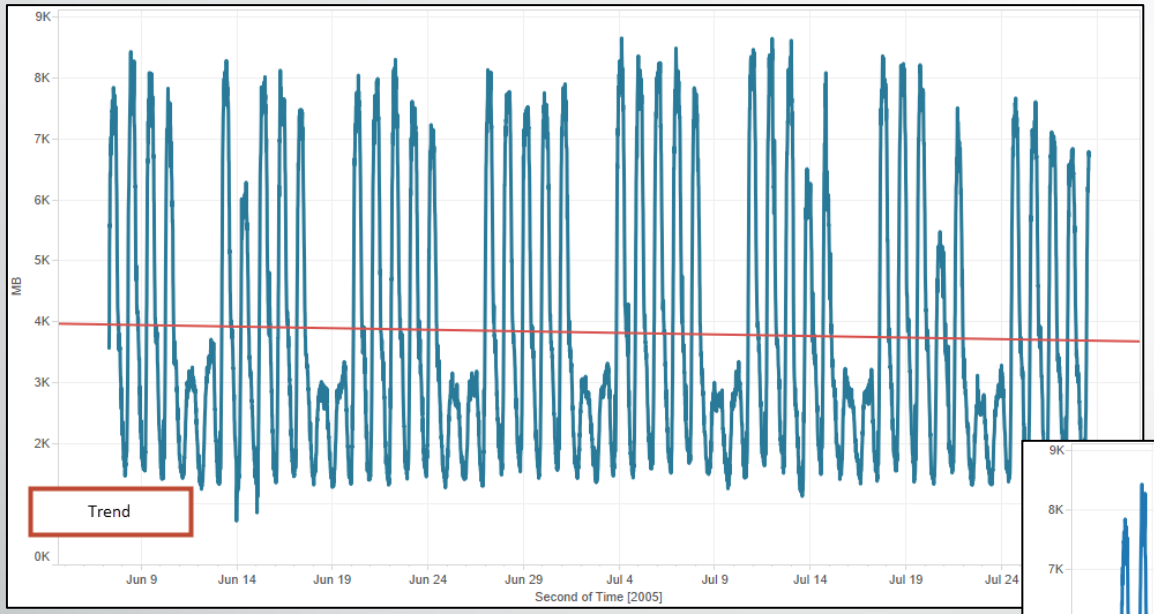
Agenda

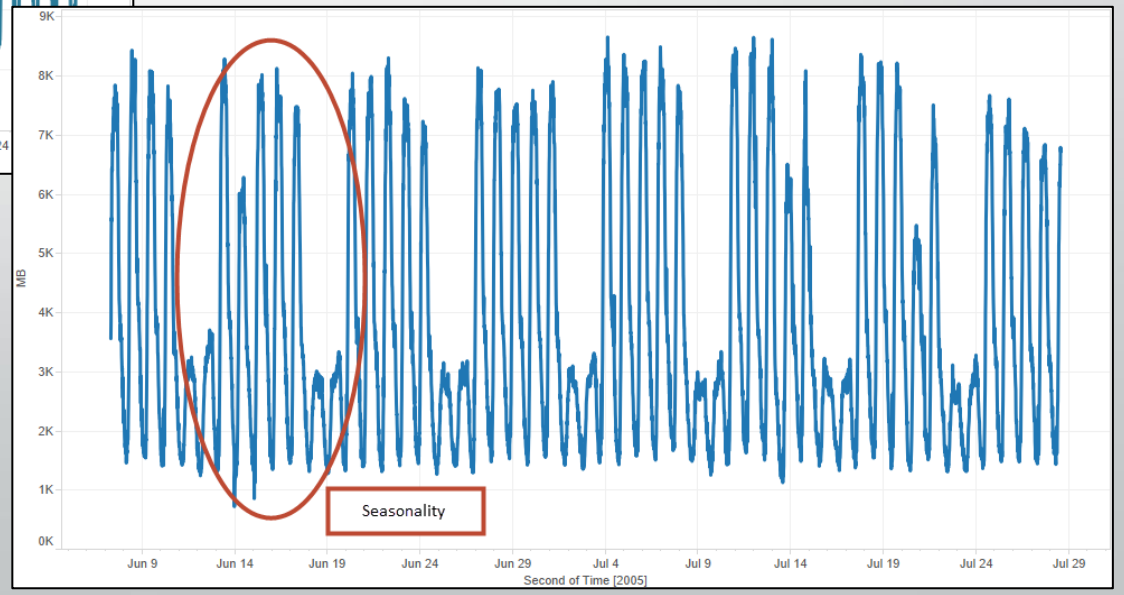
- Time Series Data
- Storage Options
- Time Series Databases
  - Open TSDB
  - MAPR
  - KDB
  - InfluxDB
  - OneTick
- Visualizing Time Series Data
  - Grafana
- Spark Time Series

Time Series Data

- Time Series Data is a set of observations collected at a usually *discrete and equally spaced time intervals*
  - General Properties
    - Trend
    - Seasonality
    - Order Matters (Values depend upon the recent observations)
- Special storage types are needed because of the volume of data and the frequency of observations

Internet Usage Data - Trend and Seasonality





Time Series Data Storage Options

Flat Files

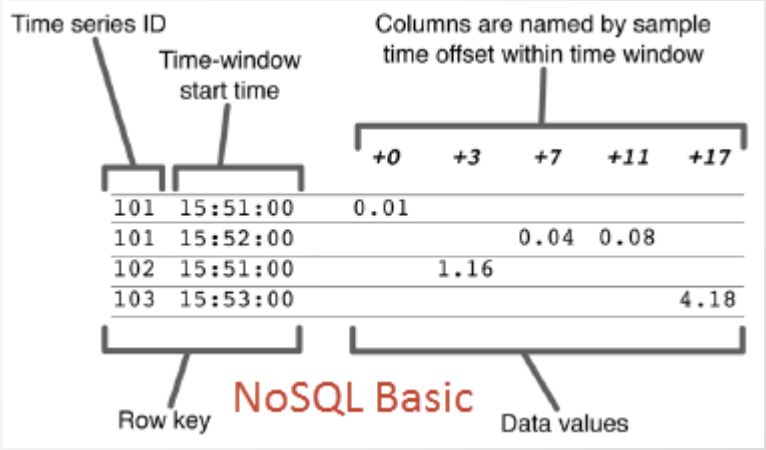
- Data will outgrow them and access is inefficient

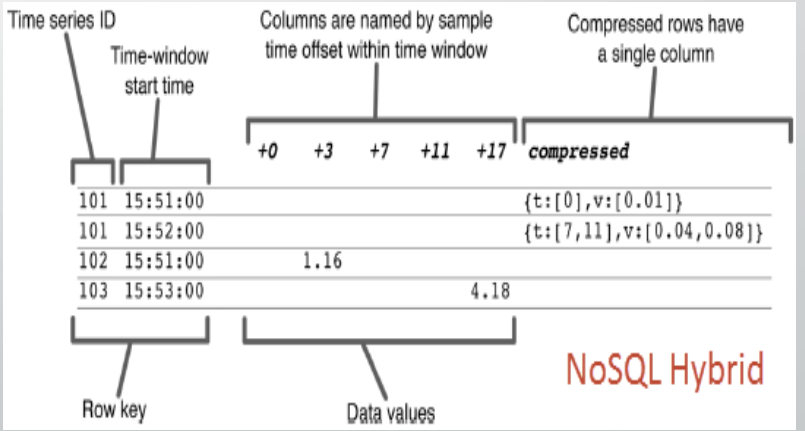
RDBMS

- Doesn't scale well

NoSQL non relational

- Preferred as it scales well
- Basic Design: Unique row keys with time series id and column is a time offset
- Hybrid Design: Blob Style





Time Series Database – Open TSDB

Time Series Database – Open TSDB

- **Time Series Daemon** (TSD)
  - Interaction with Open TSDB is achieved by running one or more TSD's
  - Independent, can run as many TSD's to handle load
  - TSD's uses open source database HBase to store and retrieve time series data
    - HBase schema is highly optimized for faster aggregations of similar time series data to minimize storage space

Time Series Database – Open TSDB

- Time Series Data Point in Open TSDB

Metric Name

- Entity that is being tracked

Time Stamp

- Time Stamp of the observation

Value

- The observation at a specific point in time

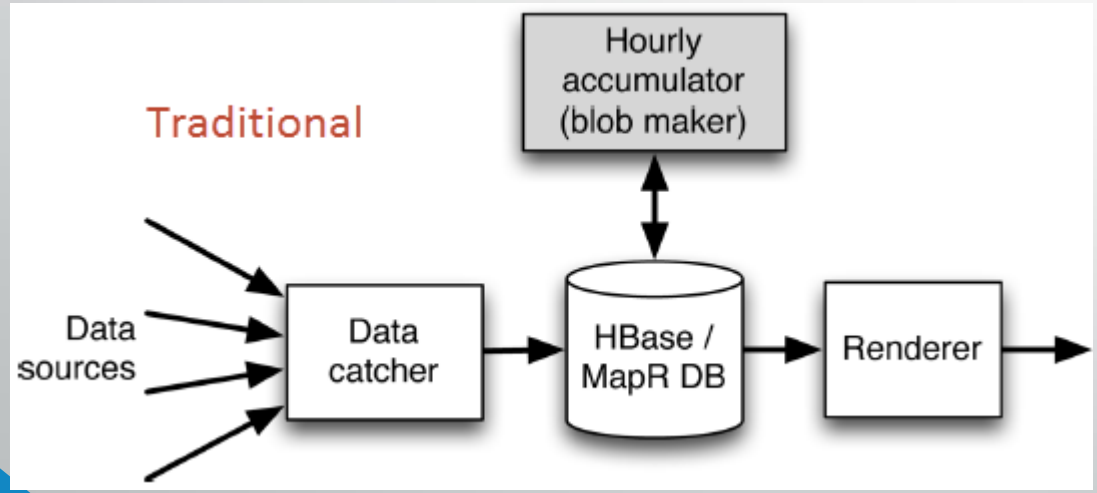
Tags

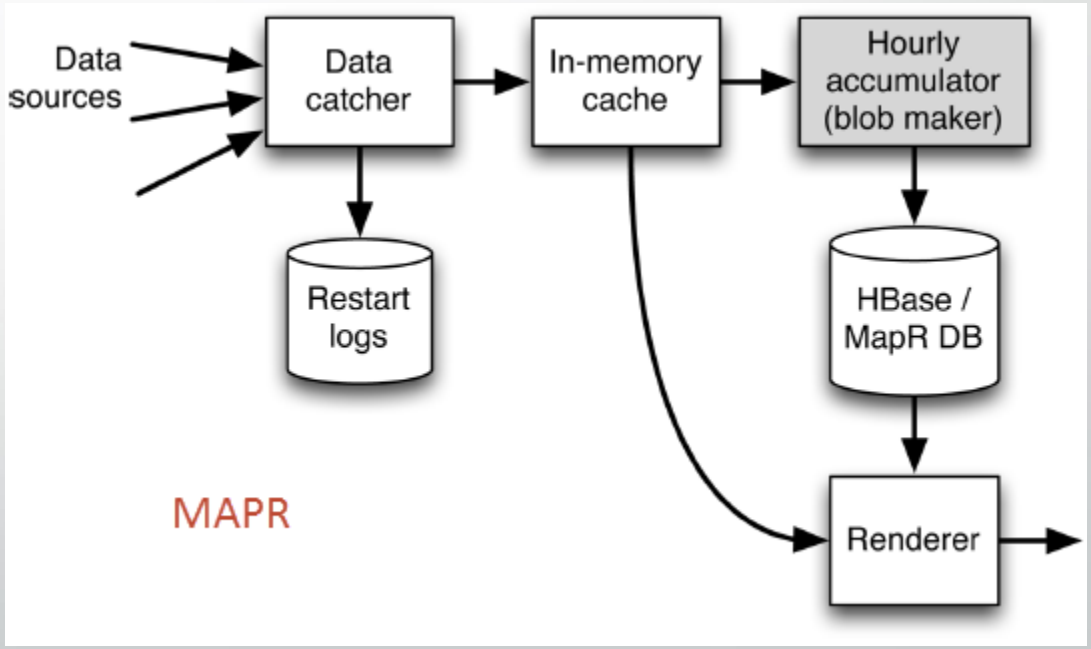
- These are used to annotate data points

Time Series Database – MAPR

Time Series Database – MAPR

- MAPR is a open source modification to extend the capabilities of Open TSDB
- Direct Blob loading
  - Avoids performance bottlenecks



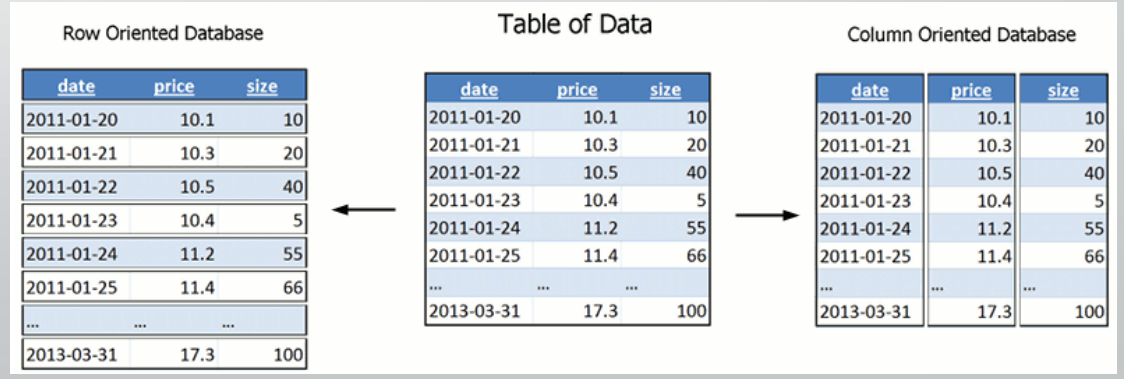


Time Series Database – KDB

Time Series Database – KDB

- KDB
  - In memory column-oriented database based on the concept of ordered lists
  - KDB uses a terse programming language Q
  - Data Store
    - RDB (Real Time Database) to store current day's data
    - HDB (Historical Database)

- Typically used in the following scenarios
  - Don't want your data outside your firm
  - Don't have in-house cloud infrastructure



Time Series Database – InfluxDB

Time Series Database – InfluxDB

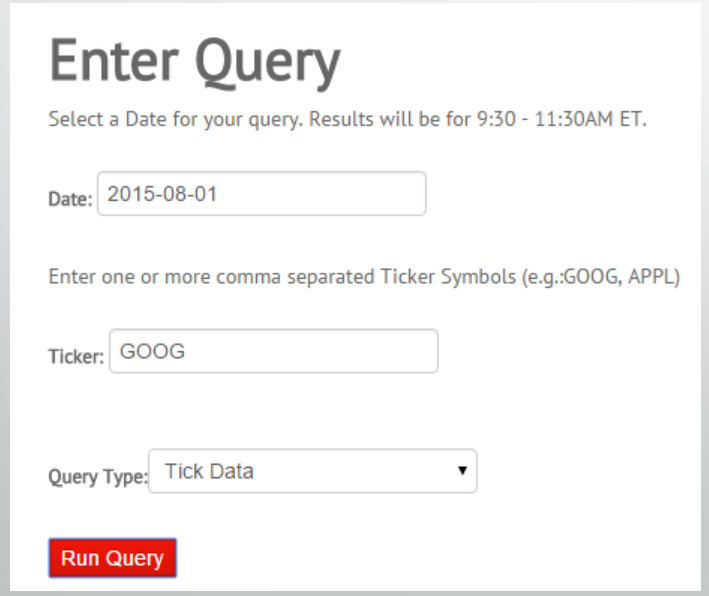
- InfluxDB open-source distributed time series database
- InfuxDB uses BoltDB as its storage engine
  - Pure Go persistence solution that saves data to memory mapped file

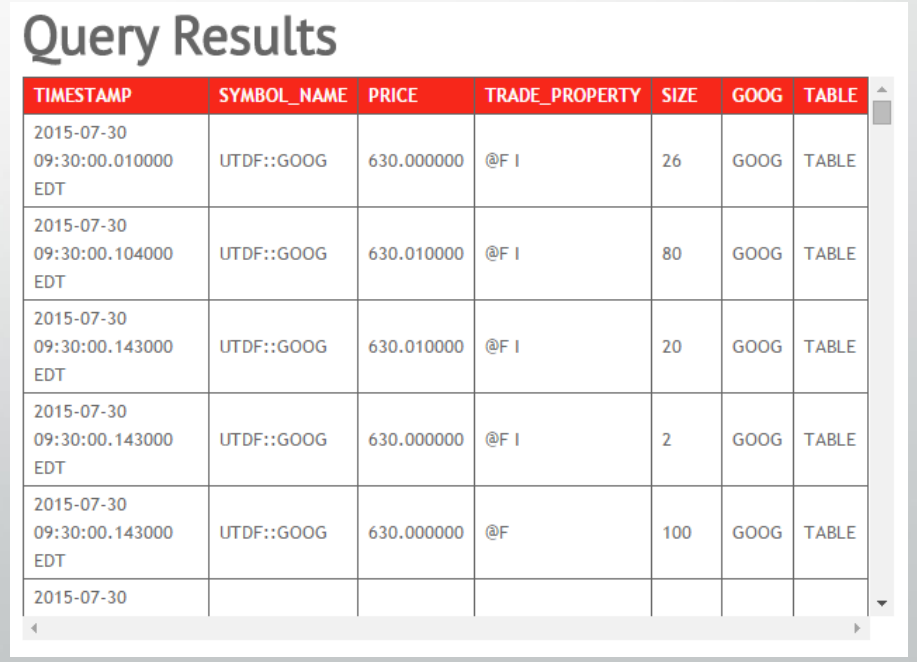
- Key Features
  - SQL – like query language
  - Database-managed retention policies for data
  - Built in management interface
  - Horizontally Scalable

Time Series Database – OneTick

Time Series Database – OneTick

- OneTickCLOUD
  - Hosted service providing managed data and analytics
  - Reference data, and adjustment factors along with on-demand analytics tools for creating custom datasets.





Visualizing Time Series Data – Grafana

Visualizing Time Series Data – Grafana

- Grafana is a open source application for visualizing large-scale measurement data.
  - Currently supports
    - Open TSDB
    - Influx DB
- Features
  - Rich Graphics
  - Dashboards
  - Templated Queries and Dashboards
  - Annotations
  - Multiple Data Sources

Spark Time Series

Spark Time Series

- **Helsinki** : *Spark.TimeSeries* is an api for Time Series Analysis in Spark
  - Depends upon *RunRDD* that splits the data into interesting partitions using a *RunDetector*
  - Partitions are represented by Runs contained *RunRDD*

- **Cloudera**: The central abstraction of the library is *TimeSeriesRDD* a lazy distributed collection of univariate series
  - Within each univariate series, observations are not distributed
  - Time Series Manipulation
    - Aligning
    - Slicing by date-time
    - Missing value imputation
  - Risk
    - Monte Carlo Simulation
    - Bootstrap Historical Simulation

Questions?