Database Solutions for Industrial Production Network

Ms. Manjiri Birajdar

Requirements

- Data volume :
  - collect many different types of IoT data that must be managed together.
  - **Example:**
    - Additive Manufacturing: 500GB a day (for each machine; 2 machines)
    - Machine tools and Robots: up to 5GB a month (for each machine tool: 8 machines)

- Variety of Data (Storage types) :
  - Time series data (timestamp; value)
  - Jpeg (with timestamp and/or spatial information) - binary large object. (blob)
  - Tif (with timestamp)
  - ASCII (with timestamp)
  - .stl and similar

Source : https://www.ibm.com/downloads/cas/G6BONV4B

Requirements

- Cost-effective data storage/processing solution
  - Which includes Infrastructure cost (Faster processors), open-source, community …
- Scalability:
  - Depending on amount of data, volume of requests and size of requests, database should scale up or down
- Performance:
  - Response time for every requests should be minimal
- Standard Protocols for Machine to Machine communication (explained in later slides)
  - Example: MQTT, OPC-UA
- Available Computing Power
  - Cloud DB OR Distributed DB (NoSQL non-relational databases) OR Time-Series Databases
  - Infrastructure, storage space, cloud or file system, connectivity among devices

… many more

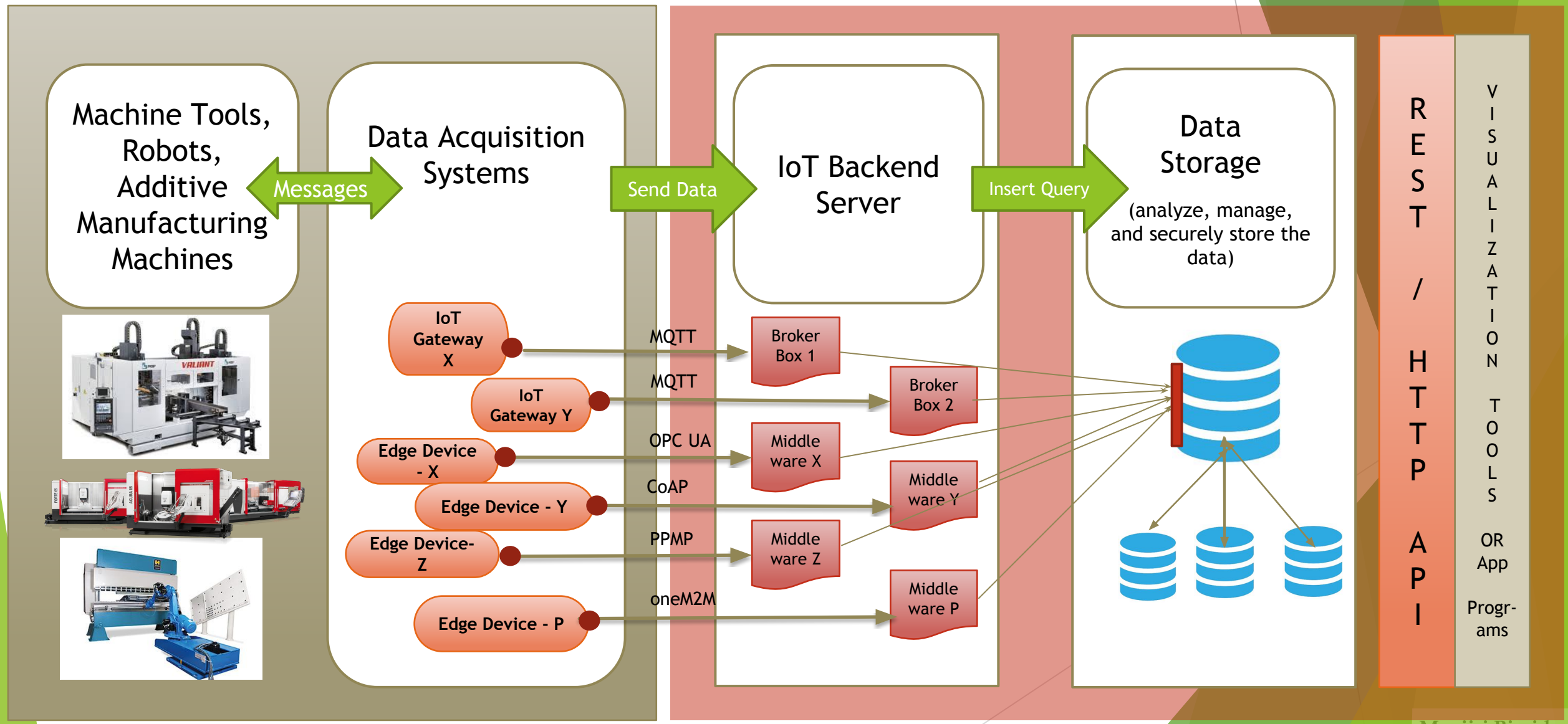
What are we going to do with our data?

- Real-time data streaming
- Time series analytics
  - to extract meaningful statistics and other characteristics of the data
- Visualization and instant analytics
- data filtering and aggregation
- near-zero latency read operations
  - no time is lost in exchange of information from one interface to another
  - the system responds instantly to an input of information.

Database Solution for IoT systems

Manjiri Birajdar

Architectural Overview



Database Overview

Manjiri Birajdar

Database Types

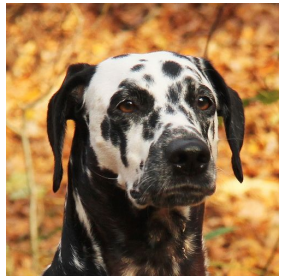
- Relational
  - table-based
  - Relational databases require to send data in a stringent (perfect) format to match the record definition in the database : Structured data
  - Example: MySQL, Oracle, PostgreSQL, and Microsoft SQL Server
- NoSQL
  - Dynamic schema
  - Unstructured data : column-oriented, document-oriented, graph-based or organized as a KeyValue store
  - Example: MongoDB, BigTable, Redis, RavenDB Cassandra, HBase, Neo4j and CouchDB
- Time Series (TSDB)
  - This process efficient than the traditional relational database
  - Can be used for handling time series data, arrays of numbers indexed by time

Why Time Series Database(TSDB) ?

- We gather Time series data
  - series of data points indexed in time order
- Analyze historical/live data to make predictions or answer research questions
- Easy Querying by timestamp
- Quick and cost-effective operations such as store, manipulate and access
- Higher number of writes
  - Data comes and goes quickly and in great number

open-source Time Series Databases

influxdata



DalmatinerDB

elasticsearch

Timescale

graphite

GridDB

Warp 10™

M3D
B

OPENTSDB

MANY MORE...

Database Comparison (in progress)

Manjiri Birajdar

Database Comaprison

DB Name → Characteristic ↓	InfluxDB	TimescaleDB
open-source time series database	Yes	Yes
A good technical documentation	Yes	Yes
visualization tool	Yes	Yes
Data volume	Large data volume	Large data volume
SQL support	Flux	SQL
Support for heavy writing load	Yes	Yes
Scalable and Reliable		Yes
Downsampling and data retention	Yes	-
Performance	High	Faster performance than Influx, Cassandra, Mongo, or vanilla PostgreSQL.
In-memory indexing	Yes	-

Communication among IoT Systems

Open source implementations for IoT protocols

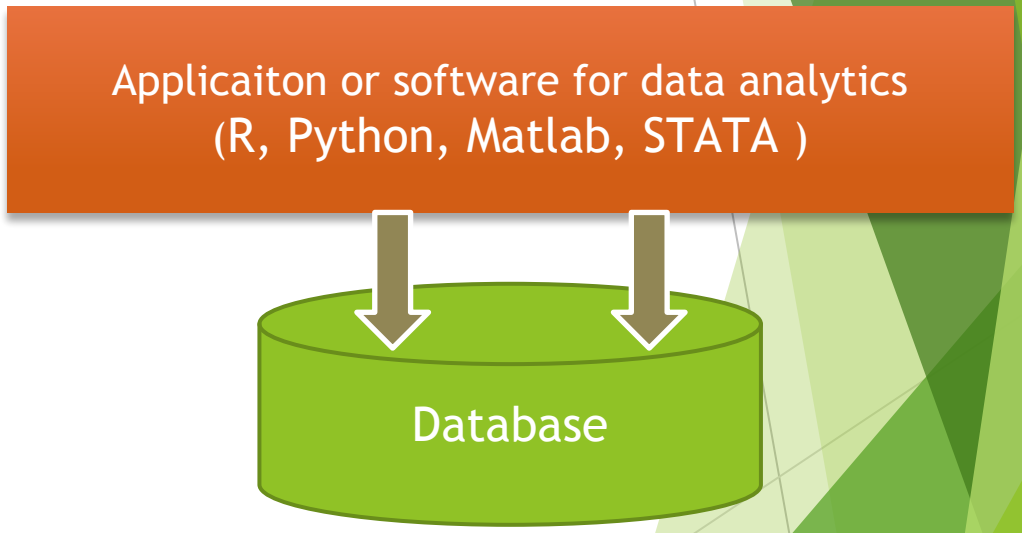
- CoAP (Constrained Application Protocol)
  - a specialized Internet Application Protocol for constrained devices
  - Existing implementations: CoAPthon, iCoAP, SwiftCoAP, FreeCoAP, etc
- oneM2M
  - Standards for M2M and the Internet of Things
- LWM2M
  - a device management protocol designed for sensor networks and the demands of a machine-to-machine (M2M) environment
- MQTT (Message Queuing Telemetry Transport)
  - a machine-to-machine (M2M)/"Internet of Things" connectivity protocol
  - There are already various MQTT implementations such as HiveMQ, Mosquitto
- OPC-UA (Open Platform Communications - Unified Architecture)
  - a machine to machine communication protocol for industrial automation

Source: https://iot.eclipse.org/standards/

Edge Analytics

Visualization and Analytics Tools

- On top of database, we can do time series analytics
- No matter where your data is, or what kind of database it lives in, you can bring it together with tools like **Grafana**



Grafana

open source software for time series analytics

Open questions...

- Which component does the actual data interpretation/processing?
- What happens if you lose the database?
  - High-availability, Disaster recovery
- How often do we send the data?
- How long data stays in the database?
  - Short period (data may arrive in high volume and velocity)
    - In-memory databases
  - Long period (historical data)
    - business intelligence service

Conclusion

- Time series database is most suitable as per our requirements
- As per gathered information, there are variety of open source time series DBs available
- We can use or build any application software on top of database for analytics
- In my personal opinion and considering our requirements, TimescaleDB or OpenTSDB are good options.
- But we make decision after comparing most of the time series databases
- NEXT STEP:
  - Finish remaining database comparison
  - Check Performance Benchmarks and create report
  - Breakdown architecture Solution into Independent Software Services
  - analyze their responsibilities and the data needs

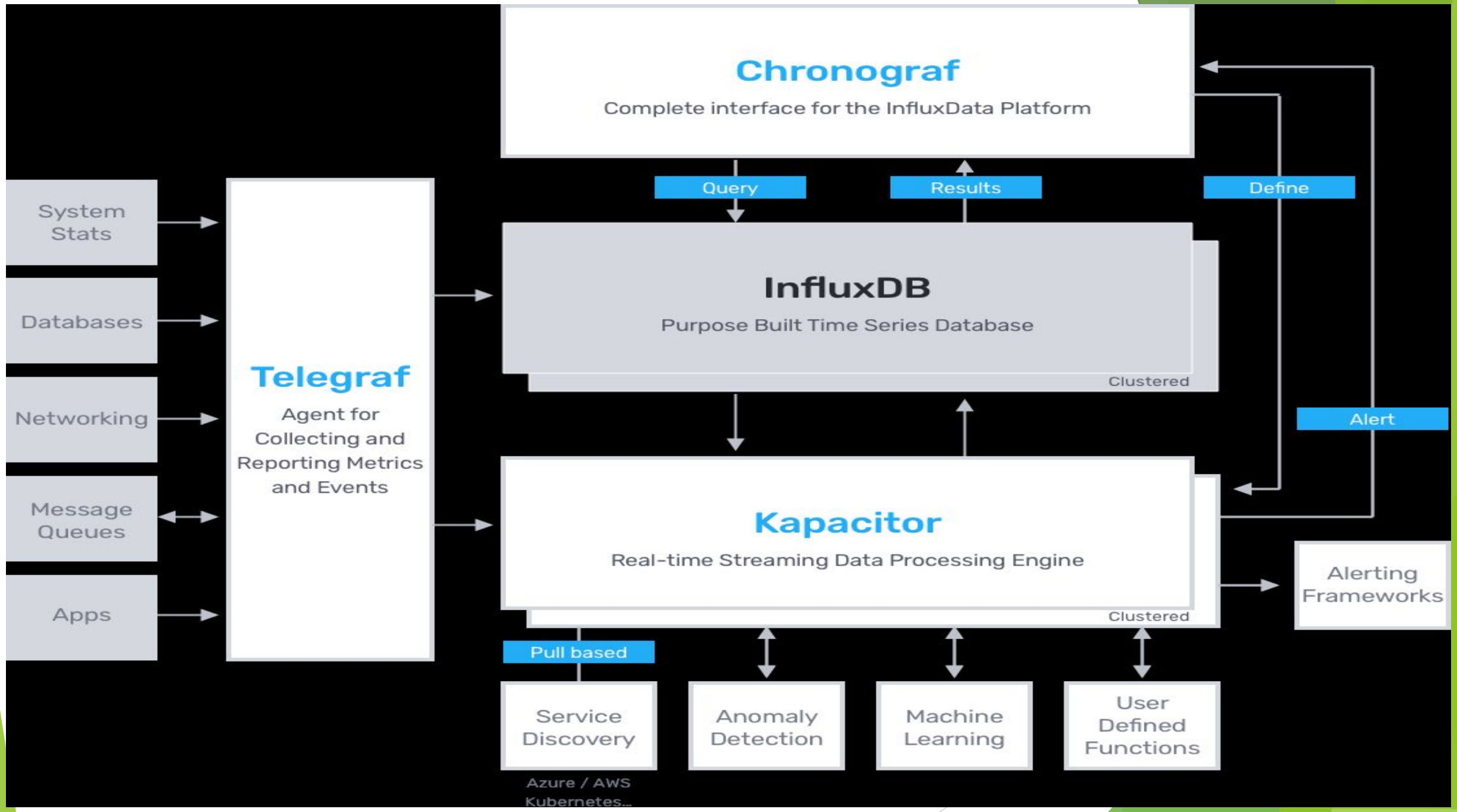
Thank you!

Requirements : Summary

- Three qualitative dimensions
  - Data model : Relational, NoSQL
  - Query language: SQL, etc
  - Reliability
- Scalability
  - performance metrics
  - high availability
  - storage capacity
- Performance benchmarks
- Database ecosystem
- Operational management
- Company/community support

Option1 : Time Series Database (**open-source**)

- InfluxDB - https://www.influxdata.com/get-influxdb/
- TimescaleDB –
  - https://docs.timescale.com/v1.3/introduction
  - https://github.com/timescale/timescaledb
- GridDB - https://griddb.net/en/
- DalmatinerDB - https://dalmatiner.readme.io/docs/getting-started
- Elasticsearch - https://www.elastic.co/downloads/
- OpenTSDB - http://opentsdb.net/
- **Graphite -** http://graphiteapp.org/#overview
- Wrap 10 by SenX - https://warp10.io
- M3DB - https://m3db.github.io/m3/m3db/
- **Object or Document DBs - MongoDB, CouchDB and Couchbase**
  - https://www.mongodb.com/what-is-mongodb


Chronograf
Complete interface for the InfluxData Platform
Query
Results
Define
System Stats
Databases
Networking
Message Queues
Apps
Telegraf
Agent for Collecting and Reporting Metrics and Events
InfluxDB
Purpose Built Time Series Database
Clustered
Alert
Kapacitor
Real-time Streaming Data Processing Engine
Clustered
Alerting Frameworks
Pull based
Service Discovery
Anomaly Detection
Machine Learning
User Defined Functions
Azure / AWS
Kubernetes...

Database Comaprison

DB Name Characteristic	OpenTSDB	GridDB
open-source time series database	Yes !	
A good technical documentation	Yes	
visualization tool	Yes	
Data volume	-	
SQL support	CLI tools, HTTP API, GnuPlot graph	
Support for heavy writing load	Yes	
Scalable and Reliable	Yes	
Downsampling and data retention	Yes	
Performance	High!	
In-memory indexing	-	

! Apache HBase™ is the Hadoop database, a distributed, scalable, big data store.

TimescaleDB vs InfluxDB

TimescaleDB	InfluxDB
Relational database (*relational data model*)	Custom, NoSQL, non-relational database (*tagset data model*)
Query language : SQL	**Query language** : Flux

Source: https://blog.timescale.com/sql-vs-flux-influxdb-query-language-time-series-database-290977a01a8a/

IoT ecosystem - five horizontal layers

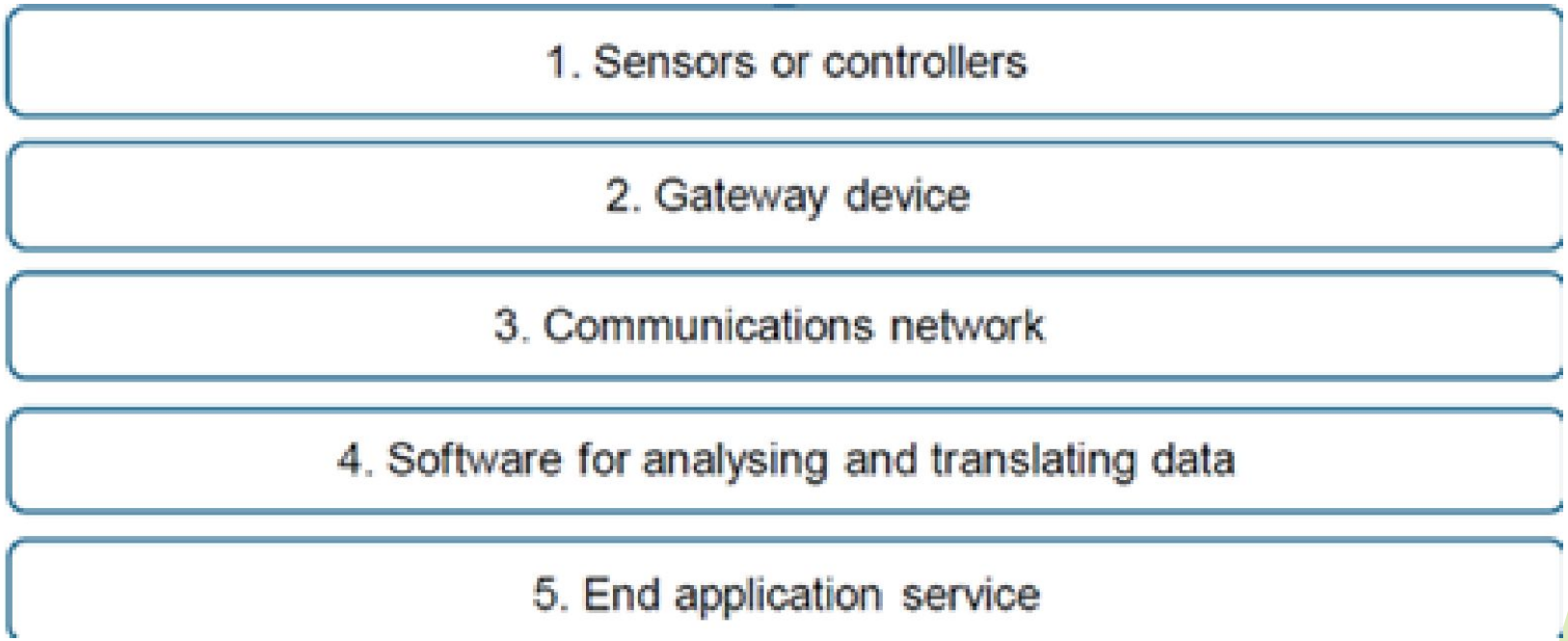


Image reference:
https://www.researchgate.net/publication/323525875_Internet_of_Things_IoT_System_Architecture_and_Technologies_White_Paper