Application of Time Series Data Bases in Business Intelligence
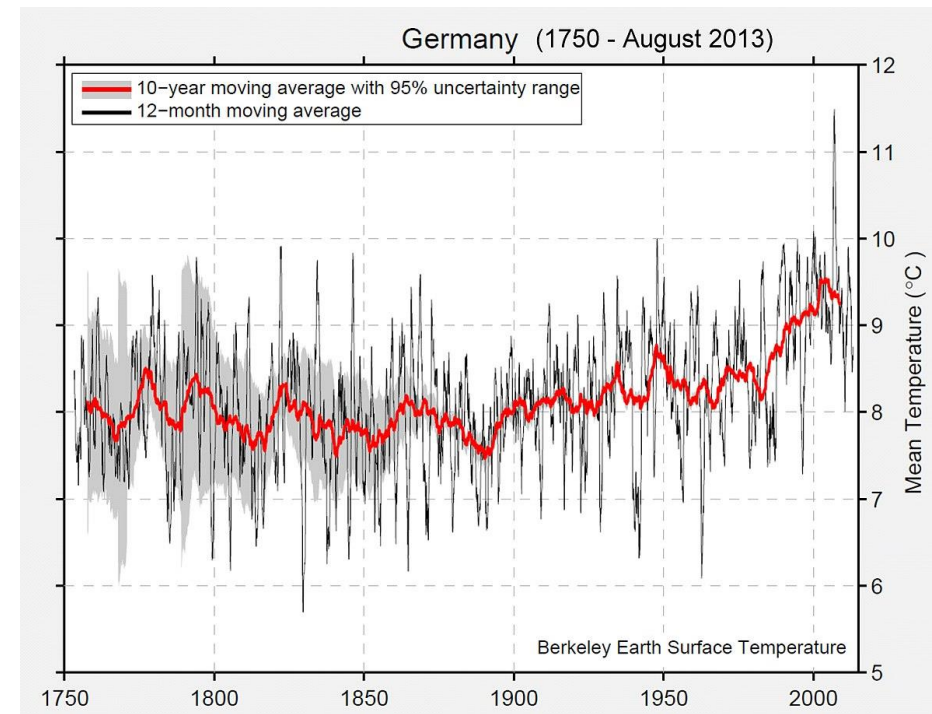
Dr. rer. nat. Sandra Geisler

Outline

- Basic principles of time series and time series data bases

- Get a glance at the InfluxDB and the TICK Stack

- Mapping the TICK Stack to a business intelligence ecosystem

- Get a glance at Druid as an OnLine Analytical Processing Time Series Data Base

Time Series Data – Be up to Date

- **Time Series:** chronologically ordered sequence of data points

- Examples: Sensor readings, stock market prices, network traffic, application health & performance data, sales information

- Challenges:
  - Data coming in with potentially high volume and frequency
  - Storage, querying, and analysis must be highly efficient
  - Scalability: ensure high availability and reliability

Dr. rer. nat. Sandra Geisler

Germany   (1750 - August 2013)

10-year moving average with 95% uncertainty range
12-month moving average

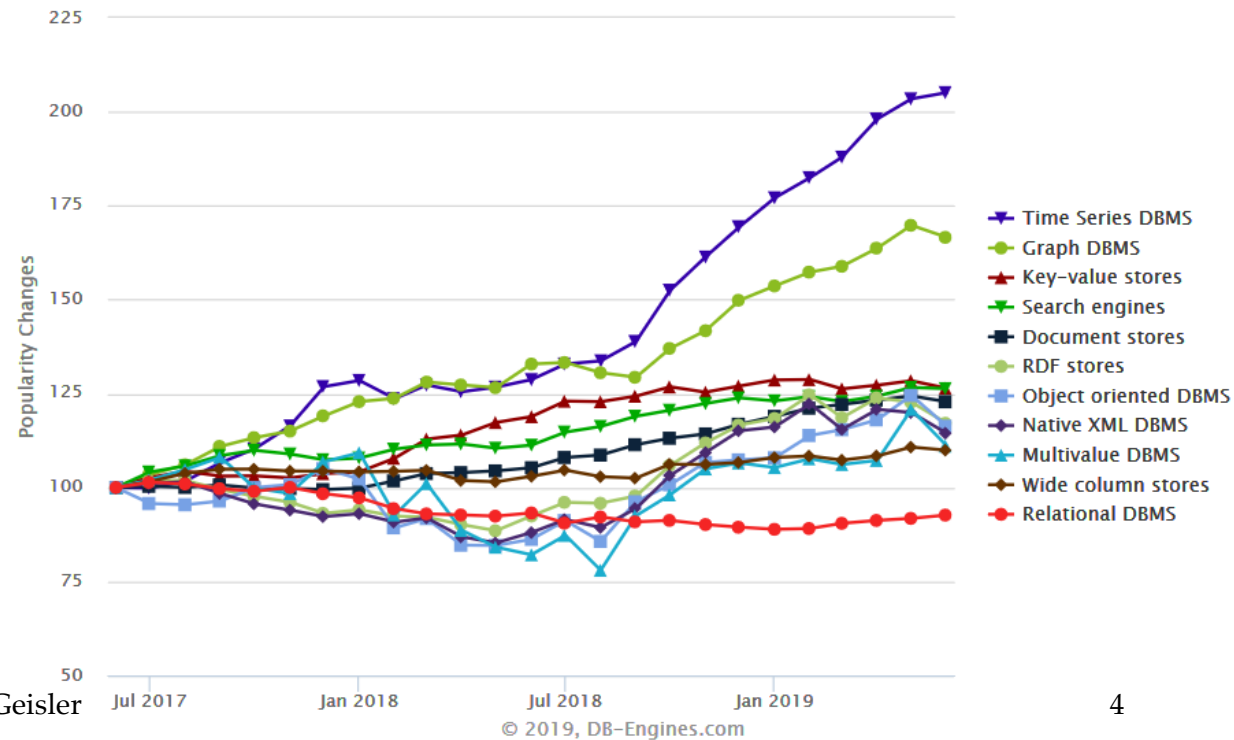Berkeley Earth Surface Temperature

Time Series Data Bases (TSDB)

- Stores time-stamped data incl. metadata (tags) and metrics (group, time-series) [1]

```
h2o_feet,location=here water_level=8.120,description="between 6 and 9 feet" 1439856000
```
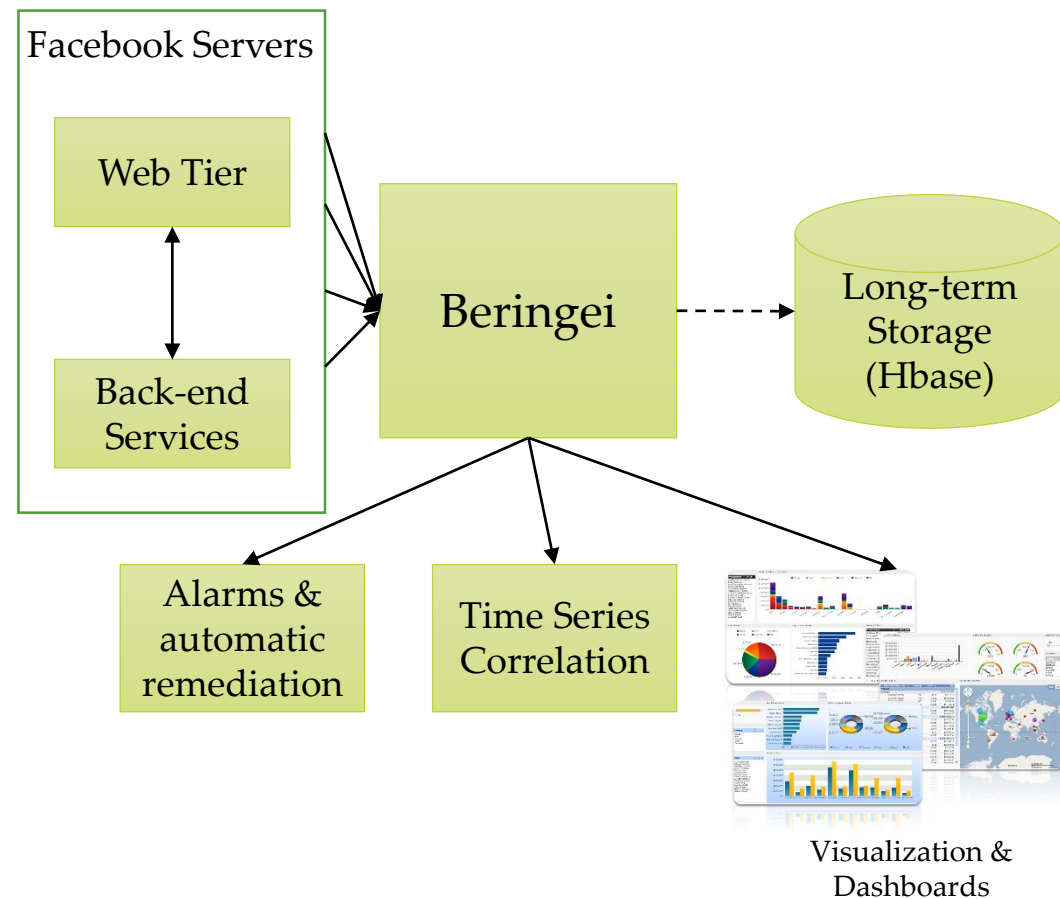
- Optimized for queries on time series
  - Aggregations over large time periods
  - Keep current data points for analysis
  - Main operations are inserts

- Storage Characteristics
  - Format: NoSQL, relational, columnar
  - Location: In-memory, disk storage
  - Distribution / Clusterability [1]
  - Depends on separate DBMS [1]
  - Granularity [1]

Trend of the last 24 months

Legend:
- Time Series DBMS
- Graph DBMS
- Key-value stores
- Search engines
- Document stores
- RDF stores
- Object oriented DBMS
- Native XML DBMS
- Multivalue DBMS
- Wide column stores
- Relational DBMS

© 2019, DB-Engines.com

Example: System Health Monitoring @Facebook [2]

- **Goal:** Health and performance monitoring to rapidly identify and diagnose server problems

- Beringei (fka Gorilla): Distributed in-memory TSDB
  - write-through cache with time series compression
  - Sharding by metric → time series can be mapped to node

- Store and query time series of system measures
  - Writes: tens of millions data points/s, 2 billion metrics
  - Storage: 26 hour storage period, granularity of 15 s, two replicas
  - Query: response times of milliseconds, 40.000 queries/second

Facebook Servers

Web Tier

Back-end Services

Beringei

Long-term Storage (Hbase)

Alarms & automatic remediation

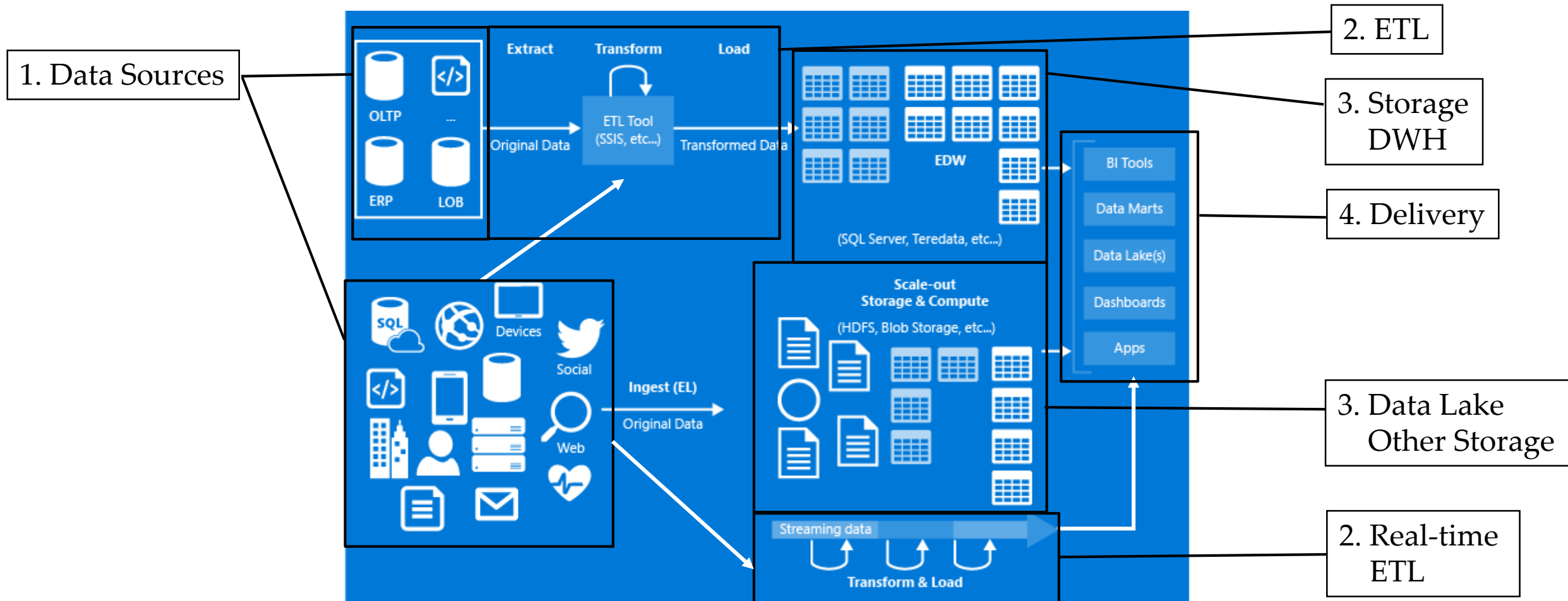Time Series Correlation

Visualization & Dashboards

Applications of TSDB in Business Intelligence

- **Business Intelligence:** Analyze company data and gain insights to improve business processes and decisions

- TSDB: provide historical data, but also near real-time data

- Combination allows near real-time predictions, pattern recognition, outlier detection, monitoring & alerting

- Example BI Applications
  - Predictive maintenance in production processes
  - Risk analysis and prediction for insurances or loans
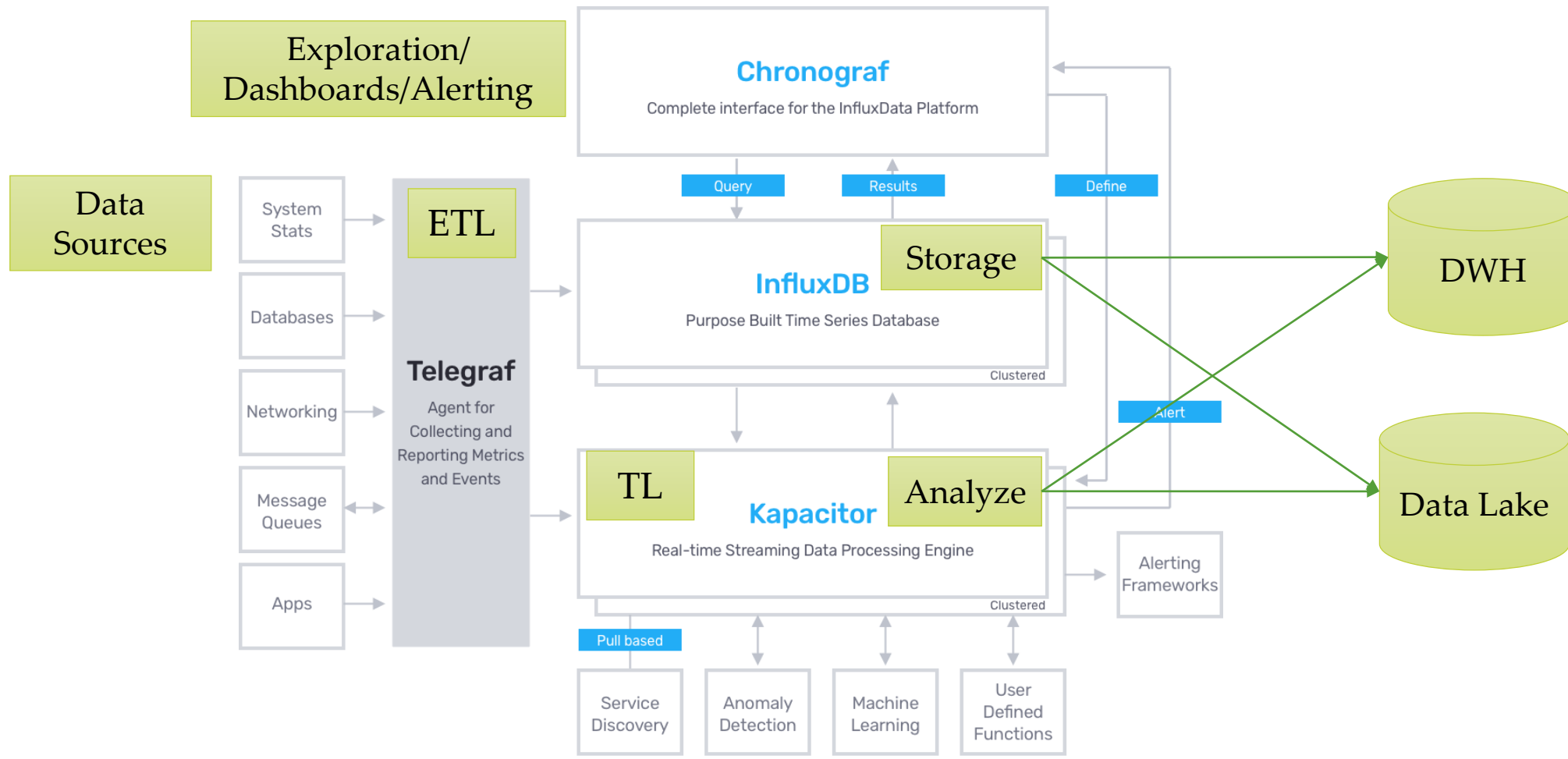 - Stock management based on real-time sales numbers

A General Business Intelligence Ecosystem

1. Data Sources

2. ETL

3. Storage DWH

4. Delivery

3. Data Lake Other Storage

2. Real-time ETL

Adapted from Source: http://www.microsofttrends.com/2014/10/29/process-your-big-data-using-new-azure-data-factory/
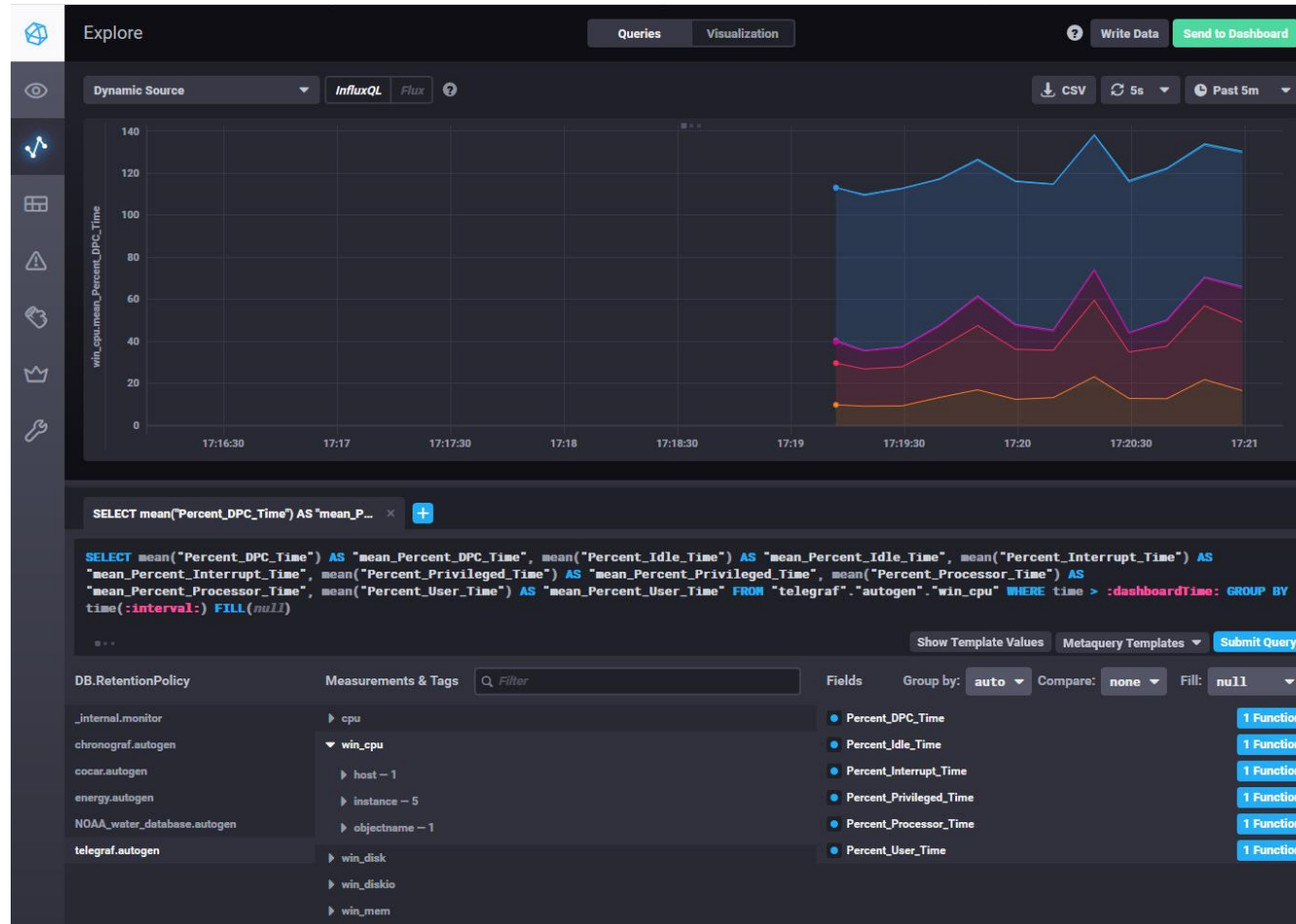
Example: InfluxDB – The TICK Stack Architecture

Demo: InfluxDB, Telegraf, Chronograf

**Example:**

System Parameter Monitoring

Example Apache Druid – Time Series OLAP [3]

- **OnLine Analytical Processing (OLAP):** data is stored pre-aggregated in a multi-dimensional schema

- **Apache Druid:** distributed column-oriented TSDB enabling slice-and-dice-analytics

- Data is partitioned by time → fast time series queries

- Millions of data points / s, can store years of data

- Used by AirBnB, Netflix, eBay, PayPal…
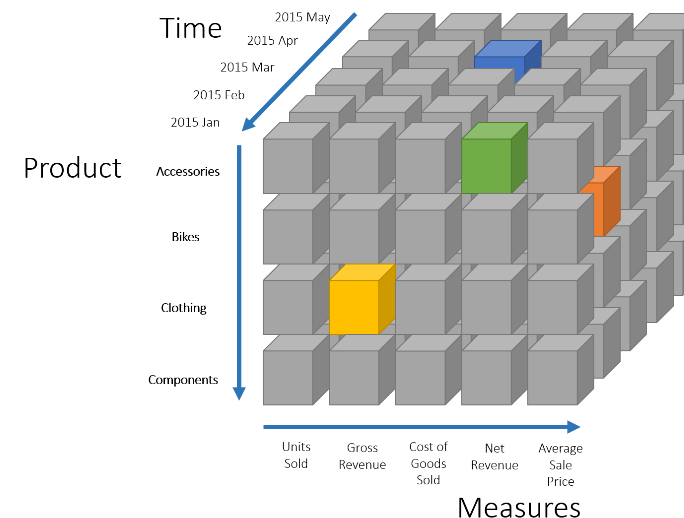
Timestamp		Dimensions				Metrics	
Timestamp		**Page**	**Username**	**Gender**	**City**	**Characters Added**	**Characters Removed**
2011-01-01T01:00:00Z		Justin Bieber	Boxer	Male	San Francisco	1800	25
2011-01-01T01:00:00Z		Justin Bieber	Reach	Male	Waterloo	2912	42
2011-01-01T02:00:00Z		Ke$ha	Helz	Male	Calgary	1953	17
2011-01-01T02:00:00Z		Ke$ha	Xeno	Male	Taiyuan	3194	170

Dr. rer. nat. Sandra Geisler

Source: http://druid.io/docs/latest/design/segments.html

Apache Superset BI Web Application
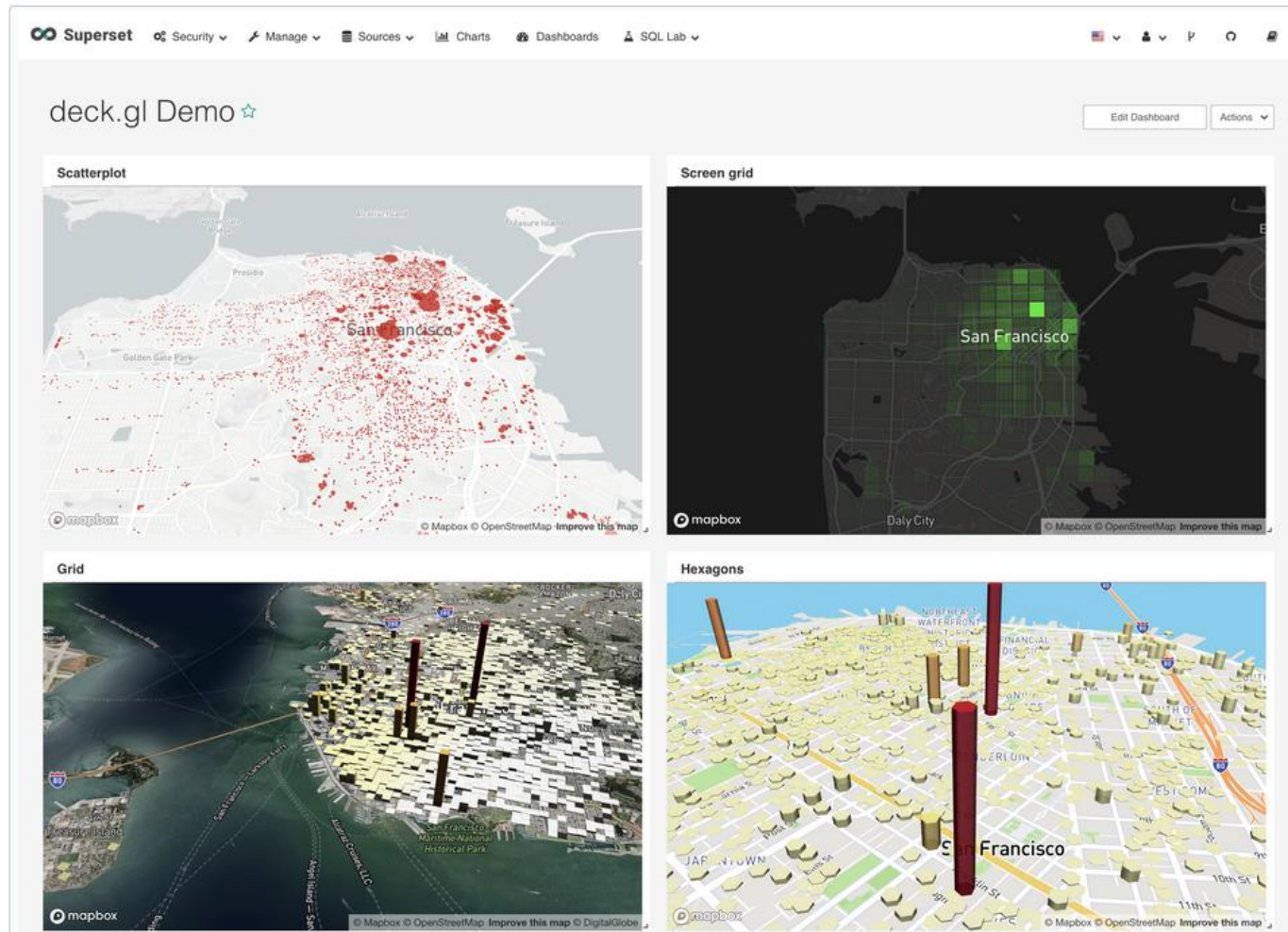
Bibliography

[1] Bader, A., Kopp, O., & Falkenthal, M. (2017). Survey and comparison of open source time series databases. *Datenbanksysteme für Business, Technologie und Web (BTW 2017)-Workshopband*.

[2] Pelkonen, T., Franklin, S., Teller, J., Cavallaro, P., Huang, Q., Meza, J., & Veeraraghavan, K. (2015). Gorilla: A fast, scalable, in-memory time series database. *Proceedings of the VLDB Endowment*, *8*(12), 1816-1827.

[3] Apache Druid (part 1): A Scalable Timeseries OLAP Database System, Anastasios Skarlatidis, https://anskarl.github.io/post/2019/druid-part-1, Last Access: 04.06.2019