Citadel: Data Hub Architecture for City Data

Jason Koh*[1], Sandeep Singh*[2], Bharathan Balaji[2], Daniel Crawl[3],
Ilkay Altintas[3], Rajesh Gupta[1], Mani Srivastava[2]
UCSD[1], UCLA[2], SDSC[3] (*equally contributed)

Source code, data, tools:
https://github.com/MetroInsight

Why Focus on Cities?

- 82% of the US population lives in urban areas.
- Cities provide essential infrastructure and services.

- Huge amount of real-time data is being generated by each sector.
- We need an End-to-End Knowledge Discovery Cyberinfrastructure for effective analysis and policy support.

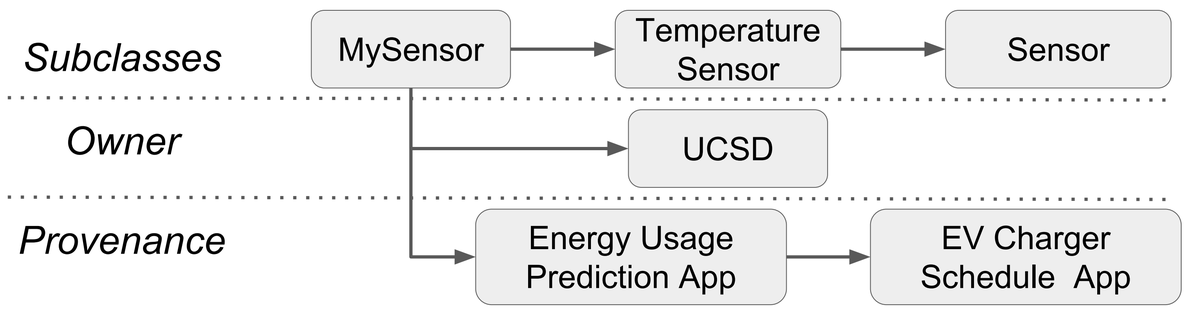
Design Goals

- ***Real time interface***
- ***Online data analytics***
- ***Usable discovery mechanism***
- ***Secured access control***

Citadel Metadata/Data model

Graph-structured Metadata, *RDF*

- Numerous different hierarchical information.
  - Types of sensors: Temperature? Humidity? Traffic?
  - Target entity of sensors: Outside air? Indoor? Car?
  - License of the data: Open-source? Closed-source?
- Many components are interrelated. It can be expressed in **RDF** easily. (nodes and directed edges)

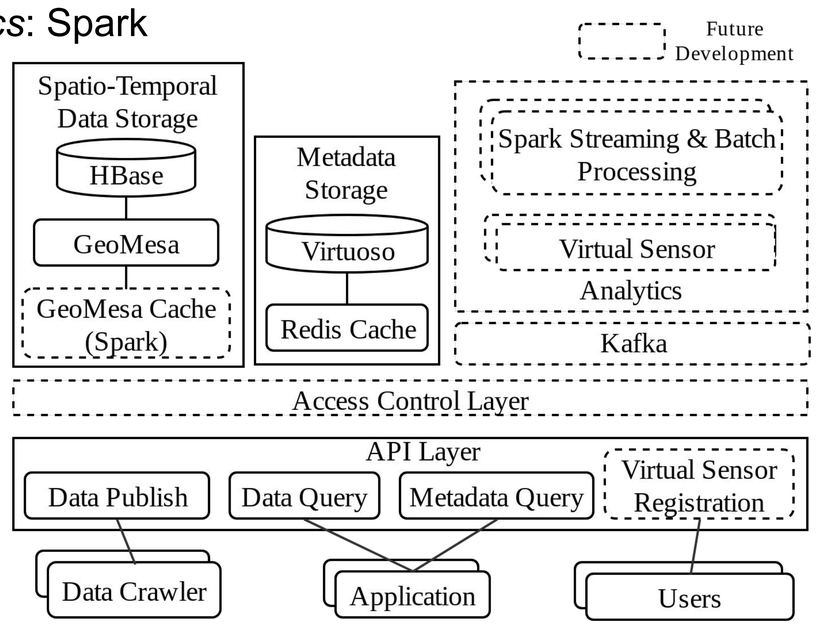


Spatio-Temporal Data

- Both time and space indices are important
  - How fast is this bus moving?
  - How long/large is this protest?
- Data Model:
  - A data point associated with spatio-temporal indices and values.
  - Spatial index can be a point, polygon, line.
  - Values are numbers currently.
  - A cell is (*Data UUID*, *Timestamp, GeometryType, List of Lat/Lng, Value)*
- **Geomesa**: GeoHash and timestamp for index + load balancing

Architecture

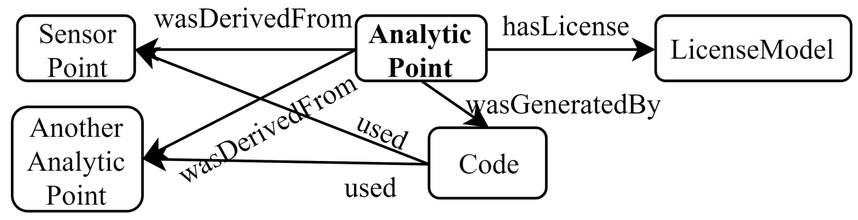
- *Spatio-Temporal DB*: Geomesa on top of HBase
- *Graph-structured Metadata DB*: Virtuoso with SPARQL.
- *REST Framework*: Vert.x for Microservice Framework.
- *Message Bus* across different Components: Kafka
- *Analytics*: Spark



- Source code, data, tools: https://github.com/MetroInsight

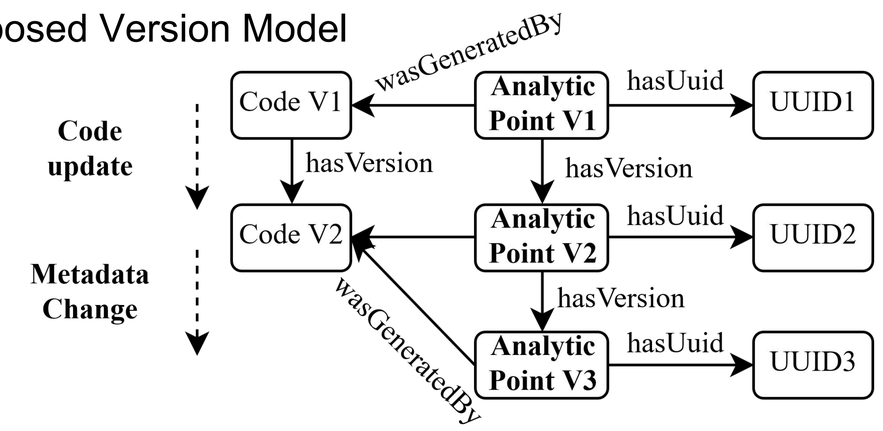
Provenance

- Users need provenance of analytics.
  - How is this analytics generated?
  - What data does this analytics refer?
  - How much credible is this result?
- Proposed Provenance Model (using PROV-O and VOAG):



  - Enables feature inheritance such as permission levels.

- Proposed Version Model



  - Enables fine-grained access control and version tracking.

Access Control

- Many datasets are confidential and are shared with authorized entity.
- Traditional Access control mechanism like Role-Based Access Control and Attribute-Based Access Control exists but are not exploited in city data hub scenario.
- General access control framework over complex data structure of heterogeneous datasets is complicated.
- In order to allow collaboration for confidential datasets, we propose a framework where access control is based on;
  - Content of data streams.
  - Attributes of data streams with focus on space, time and values.
  - Allow controlling resolution of data stream.
  - Expressing data sharing policies and exploring policy enforcement.

Preliminary Evaluation

Metadata Management Performance

Current metadata types per point: *name, point type, unit*
Configuration: In Azure cloud, a host machine (8 cores, 32GB), 8 clients in parallel. Averaged over 10,000 requests.
Data: Randomly generated, 100,000 points, 300 types, 50 unites.

Spatio-Temporal DB Performance

Current data types per point: *latitude, longitude, timestamp, num value*
Configuration: In Azure cloud, a host machine (8 cores, 32GB), single client averaged over 1000 requests.
Data: 1,000,000 points randomly generated over a region and a time period.

Metadata	
Benchmark Name	**Latency (ms)**
Point Creation	55
Query by Point Name	51
Query by Point Type	51

Spatio-Temporal Data	
Benchmark Name	**Latency**
Data Ingestion	330 sec / 729K point
Query a Bounding Box	200 ms per query (fetching 1K points from 1M points)

Acknowledgements

This research is funded in part by the National Science Foundation under awards IIS-1636916, IIS-1636879, IIS-1636936, OAC-1640813, CI-1331615, and CSR-1526841, and by the King Abdullah University of Science and Technology under KAUST Sensor Initiative.