

Data Ingestion at Scale



Jeffrey Sica
ARC-TS
@jeefy

Overview

What is Data Ingestion?

Concepts

Use Cases

- GPS collection with mobile devices

- Collecting WiFi data from WAPs

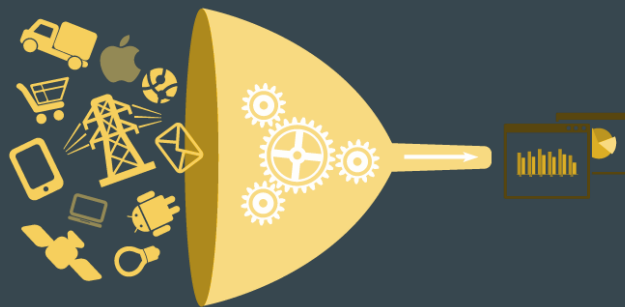
- Sensor data from manufacturing machines

Other Technologies

Questions

What is Data Ingestion?

“Data ingestion is the process of obtaining and importing data for immediate use or storage in a database. To ingest something is to "take something in or absorb something." Data can be streamed in real time or ingested in batches.”



<http://whatis.techtarget.com/definition/data-ingestion>

Concepts

Data Emitter / Generator

Ingest Point

Queue / Message Bus

Processing

Add To Datastore



Use Cases

Opt-In mobile app collects GPS coordinates to analyze mobility patterns on campus

Polling campus wireless access points to generate path/collision data (and analyze mobility patterns on campus)

Capture an array of sensor data on manufacturing equipment to generate more accurate quality models (and predict failures)

Use Case - GPS Collection

“The RITMO project is funded by the Michigan Institute of Data Science and aims at reinventing urban transportation and mobility.”

Michigan App collects GPS Data (Opt-In)

<https://ritmo.engin.umich.edu/mobility-app/>

Eventually aggregating with Wi-Fi data to further identify mobility patterns

Led by Pascal Van Hentenryck

<https://ritmo.engin.umich.edu/>

Use Case - GPS Collection

Requirements:

- Filter any erroneous (or malicious) data

 - Verify source is mobile

 - Discard large jumps in position

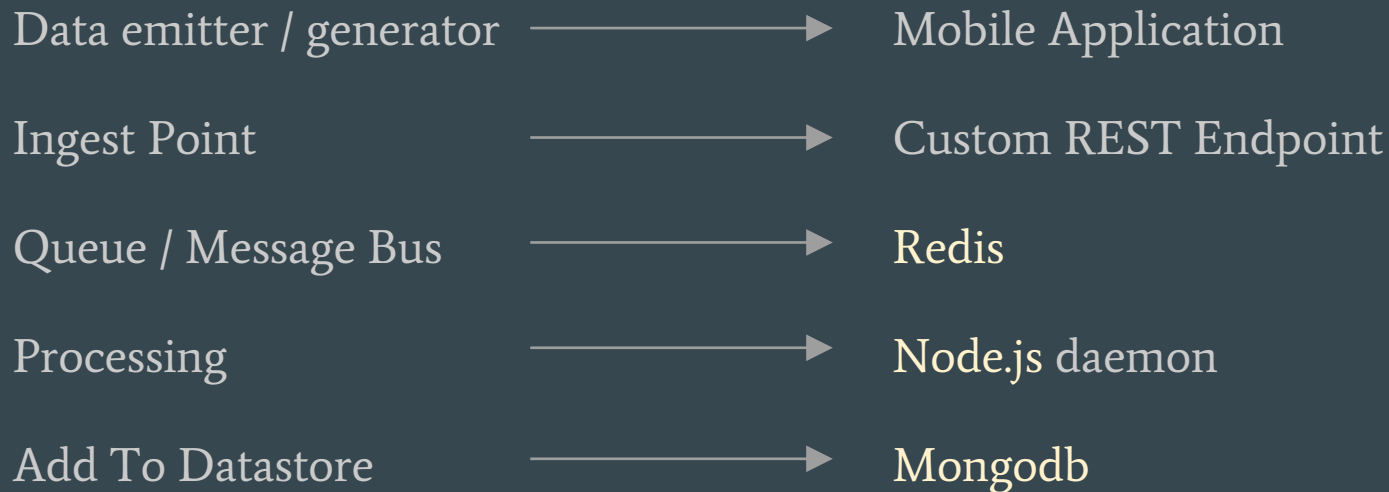
Scale

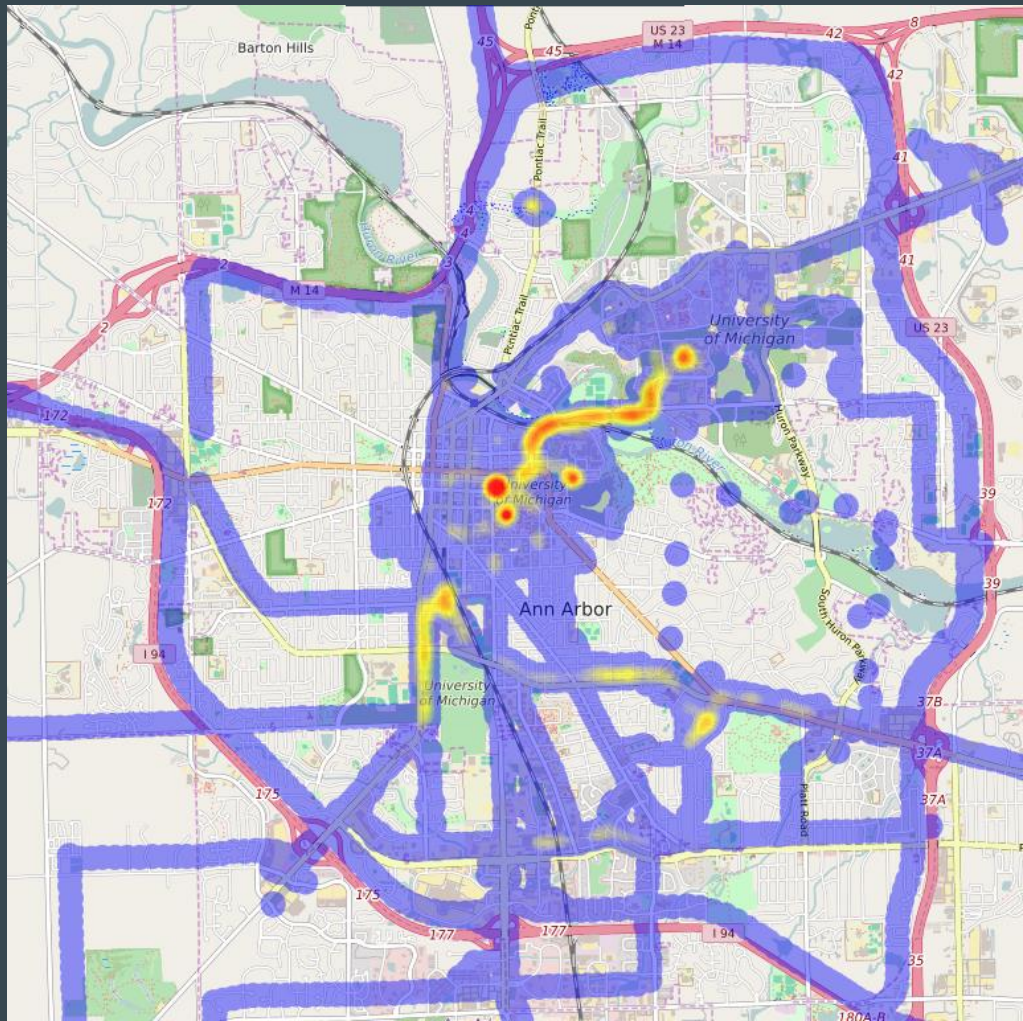
- Upwards of 50,000 potential clients

Iterative development

- CI/CD System

Use Case - GPS Collection





Use Case - Collecting WiFi Data

Privascope Initiative

Service for UM faculty to perform aggregate queries against sensitive data

Privacy-centric

Algorithms run are approved

Results released in two-step process (Automated, then review board)

Programming language agnostic

Data format / service agnostic (mostly)

Project led by **Eric Boyd** (Director of Research Networks, ITS)

Use Case - Collecting WiFi Data

Asking the right (aggregate) questions

Inappropriate

Where did my girlfriend/boyfriend go yesterday?

What are your grades? How often do you go to the gym?

What is the normal movement pattern of Professor X at lunchtime?

Appropriate

If 26% of engineers are female, what is the % of female students on North Campus at noon? What is the % at 3 AM? Do we have problem of perceived safety impacting female engineering students' ability to fully participate in academic endeavors?

Does going to the gym for at least an hour a day impact grades? How?

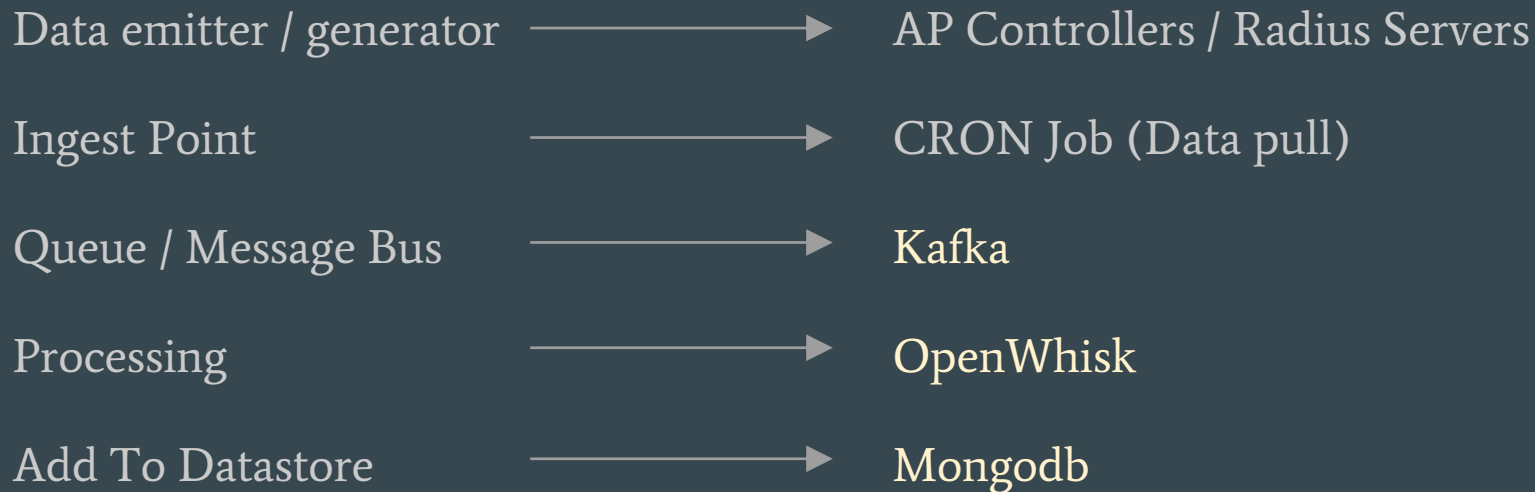
Do we see patterns (“cowpaths”) of personal movement for large groups of people?

Use Case - Collecting WiFi Data

Example WiFi Data

ericboyd	umich.edu	2017-01-17 10:22:10	2017-01-17 10:40:34	2017-01-17 10:40:34	ARBL2-2236	MWireless	D8:BB:2C:24:0D:57	35.2.86.115	fe80::/64	
ericboyd	umich.edu	2017-01-17 10:22:10	2017-01-17 10:35:12		ARBL2-2236	MWireless	D8:BB:2C:24:0D:57	35.2.86.115	fe80::/64	
ericboyd	umich.edu	2017-01-17 10:22:10	2017-01-17 10:33:35		ARBL2-2236	MWireless	D8:BB:2C:24:0D:57	35.2.86.115	fe80::/64	
ericboyd	umich.edu	2017-01-17 10:22:10	2017-01-17 10:22:10		ARBL2-2236	MWireless	D8:BB:2C:24:0D:57	35.2.86.115		
ericboyd	umich.edu	2017-01-17 10:14:11	2017-01-17 10:19:40	2017-01-17 10:19:40	ARBL3-1001-N	MWireless	D8:BB:2C:24:0D:57	35.2.86.115	fe80::/64	
ericboyd	umich.edu	2017-01-17 10:14:11	2017-01-17 10:14:55		ARBL2-2236	MWireless	D8:BB:2C:24:0D:57	35.2.86.115	fe80::/64	
ericboyd	umich.edu	2017-01-17 10:14:11	2017-01-17 10:14:41		ARBL2-2236	MWireless	D8:BB:2C:24:0D:57	35.2.86.115	fe80::/64	
ericboyd	umich.edu	2017-01-17 10:14:11	2017-01-17 10:14:37		ARBL2-2236	MWireless	D8:BB:2C:24:0D:57	35.2.86.115	fe80::/64	
ericboyd	umich.edu	2017-01-17 10:14:11	2017-01-17 10:14:11		ARBL2-2236	MWireless	D8:BB:2C:24:0D:57	35.2.86.115	fe80::/64	
ericboyd	umich.edu	2017-01-17 08:21:39	2017-01-17 08:22:02		540LIB-100	eduroam	C4:B3:01:AE:63:E2	35.2.120.118	fe80::/64	
ericboyd	umich.edu	2017-01-17 08:21:38	2017-01-17 09:46:51	2017-01-17 09:46:51	540LIB-100	eduroam	C4:B3:01:AE:63:E2	35.2.120.118	fe80::/64	
ericboyd	umich.edu	2017-01-17 08:21:38	2017-01-17 09:21:48		540LIB-100	eduroam	C4:B3:01:AE:63:E2	35.2.120.118	fe80::/64	
ericboyd	umich.edu	2017-01-17 08:21:38	2017-01-17 08:21:38		540LIB-111-H	eduroam	C4:B3:01:AE:63:E2	35.2.120.118		
ericboyd	umich.edu	2017-01-17 08:20:39	2017-01-17 09:48:17	2017-01-17 09:48:17	540LIB-100	MWireless	D8:BB:2C:24:0D:57	35.2.86.115	fe80::/64	
ericboyd	umich.edu	2017-01-17 08:20:39	2017-01-17 09:44:42		540LIB-111-H	MWireless	D8:BB:2C:24:0D:57	35.2.86.115	fe80::/64	
ericboyd	umich.edu	2017-01-17 08:20:38	2017-01-17 09:44:54		540LIB-100	MWireless	D8:BB:2C:24:0D:57	35.2.86.115	fe80::/64	
ericboyd	umich.edu	2017-01-17 08:20:38	2017-01-17 09:24:40		540LIB-100	MWireless	D8:BB:2C:24:0D:57	35.2.86.115	fe80::/64	

Use Case - Collecting WiFi Data



Use Case - Sensor Data

“The long-term goal of this research project is to develop methods and techniques to make large-scale manufacturing systems safer, more secure, and more productive, enabling them to produce high-quality products for consumers at modest cost.”

Collect array of sensor data from manufacturing machines

Develop “optimal output” models

Feed data on performance of created parts back into model

Ultimate Goals:

Better monitor manufacturing equipment


Identify faulty parts before put into use

Use Case - Sensor Data

Data emitter / generator  Manufacturing sensors

Ingest Point  Kafka Producer

Queue / Message Bus  Kafka Consumer

Processing  Hadoop

Add To Datastore  HDFS (Historical) / InfluxDB (Real-Time)

Other Technologies - Message Queue/Bus

RabbitMQ - Widely used open source message broker. Large range of support.

ZeroMQ - Highly performant message protocol. No central server, low latency.



Other Technologies - Data Stores

MySQL, PostgreSQL, etc. - Basic RDBMS solution

Cassandra - Highly scalable “SQL” solution

ElasticSearch - Highly scalable JSON document store

Cloud SQL (Google Spanner, Amazon RDS, Azure SQL)

Other Technologies - Stream Processing

Apache Flink - Open-source stream processing framework for distributed, high-performing, always-available, and accurate data streaming applications.

Apache Nifi - Supports powerful and scalable directed graphs of data routing, transformation, and system mediation logic

Apache Storm - Reliably process unbounded streams of data, doing for realtime processing what Hadoop did for batch processing

Spark Streams - Brings Apache Spark's language-integrated API to stream processing, letting you write streaming jobs the same way you write batch jobs.

Questions?