“Weather Forecast Evaluator”

W205 Final Project, August 10, 2016

XT 'Emitter' Nguyen

Rama
'Programming in the Cloud'
Thamman

Christopher 'Bolt' McPherson

Chuck 'Spout' Bolin

Background

Ohio native Chris McPherson said “**Enough is enough! We need better forecasting.**”

The goal of this project is simple.

- Assign each city a score corresponding to weather forecasting accuracy.
- Place this information into the hand of the consumer.
- Investigate these questions.
  - Where (cities worldwide)
  - When (seasons)
  - Which measures(rainfall, temperature)

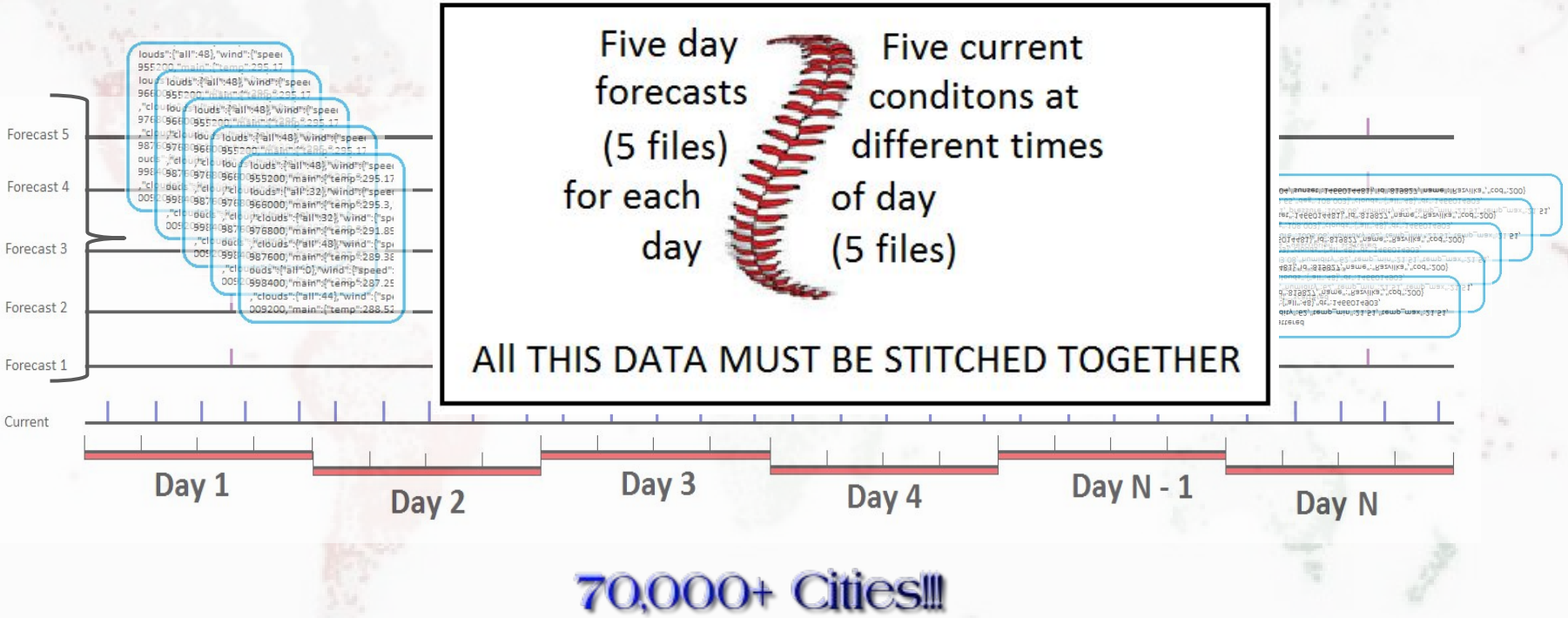
What is the *Weather Forecast Evaluator (WFE)*?

The ***WFE*** is a **weather data collection system** that stores and transforms data for the **purpose of calculating up to date forecasting accuracy scores** for over 70,000 cities worldwide.

- Allow for **ranking of forecasting accuracy** by city, region, and country.
- Individualized **scores** are provided **for each weather metric**
- Evaluate various models.
  - Identify **those that need better modeling.**
  - Recognize **superior predictive models.**
- Scores benefit industries and local governments worldwide in areas such as farming, construction, outdoor entertainment, travel industry, city maintenance, and more.

The *Weather Forecast Evaluator* Challenge

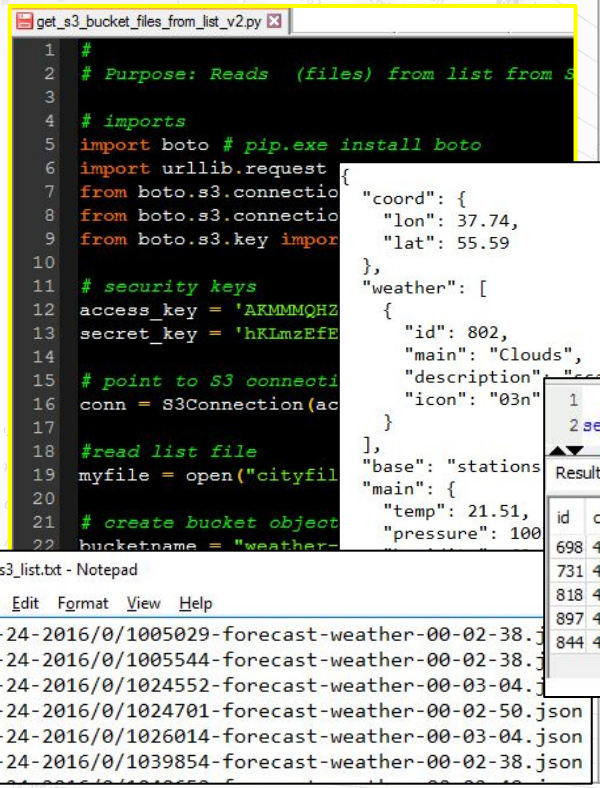
- The ***WFE*** has collected forecast and current weather data for **46 days**.



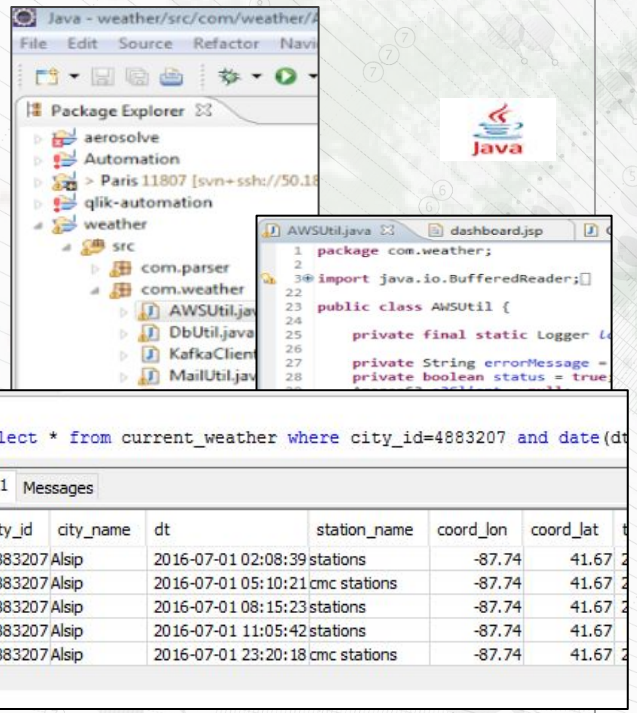
Architecture Overview

Architectural Evolution

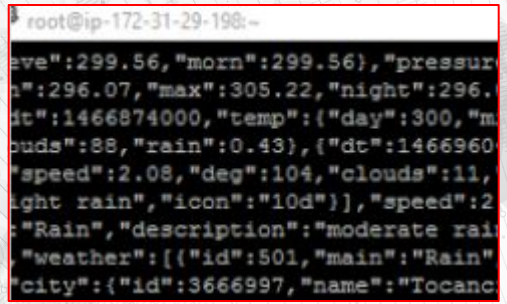
Initial API Exploration (Week 1)



"Quick and Dirty" (Week 1- 12)



Production Ready (Week 11, 12)

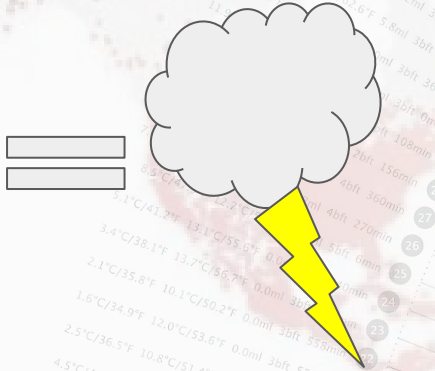


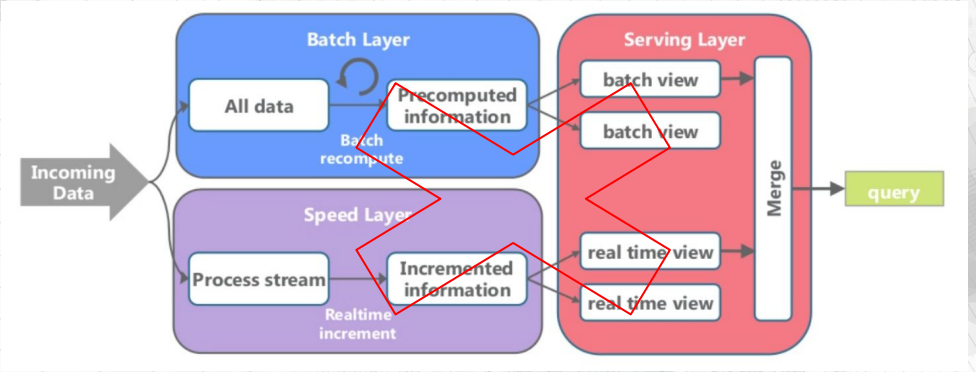
We could not sit back and wait for the production system to be created.

- Initial exploration of API using Python
- "Quick and Dirty" system to start archiving files.

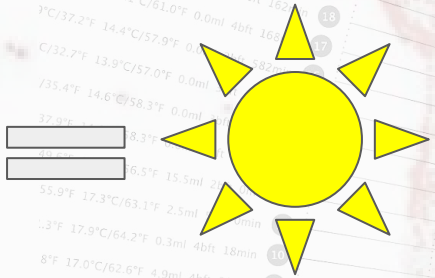
Architecture

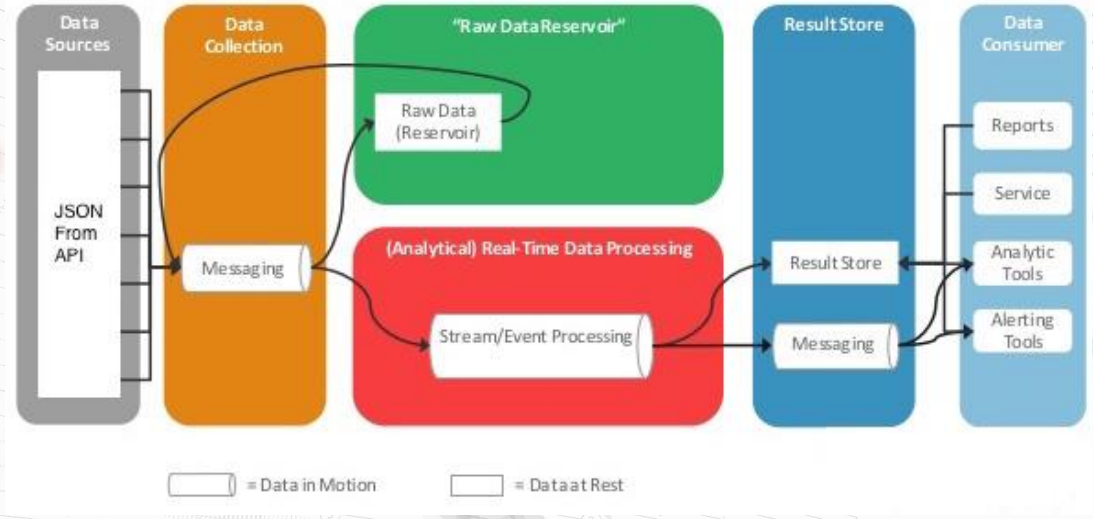
Lambda Architecture





Kappa Architecture





Tools

Storage

Application Metadata

Data Storage

Hive Metastore

Streaming

Messaging System

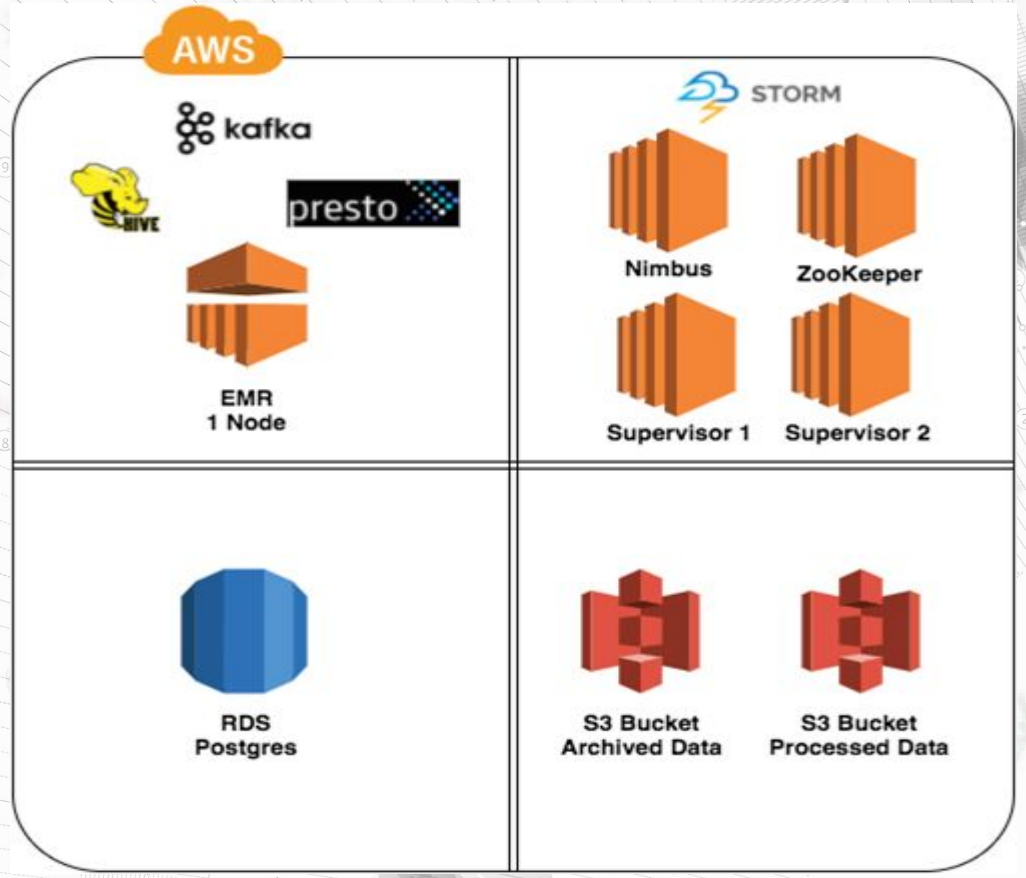
Stream Processing

Analytics

Query Engine

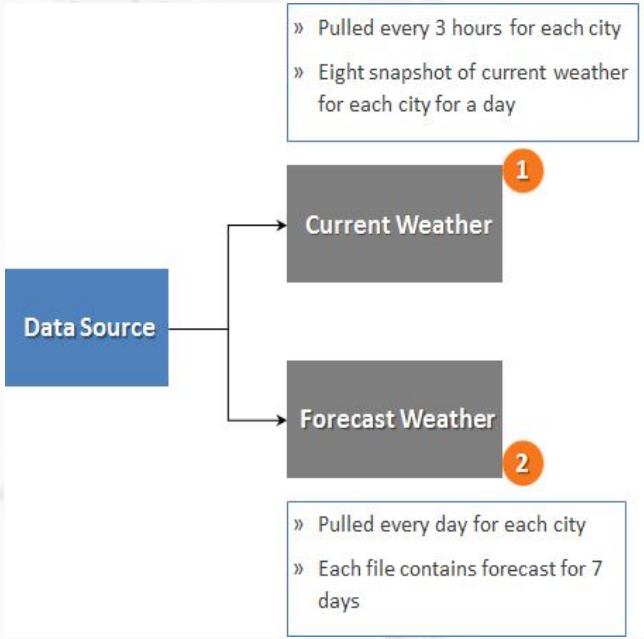
Visualizations

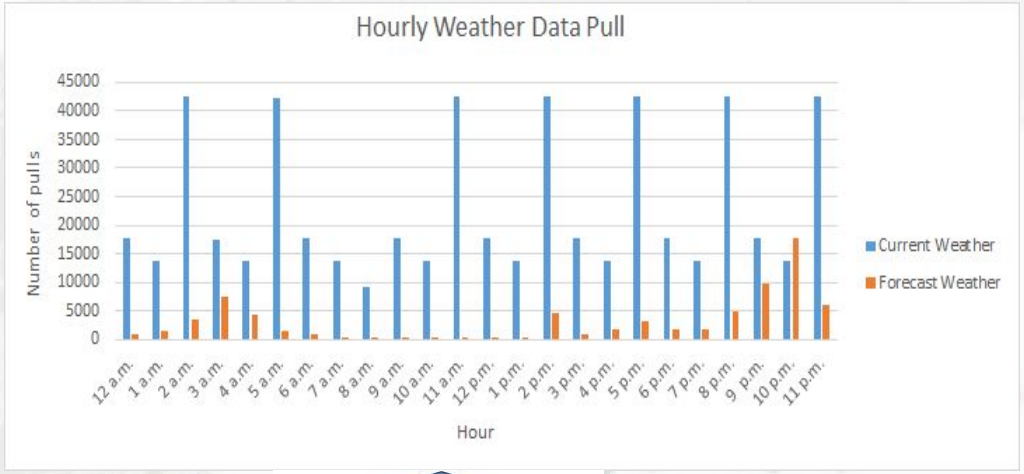
Infrastructure

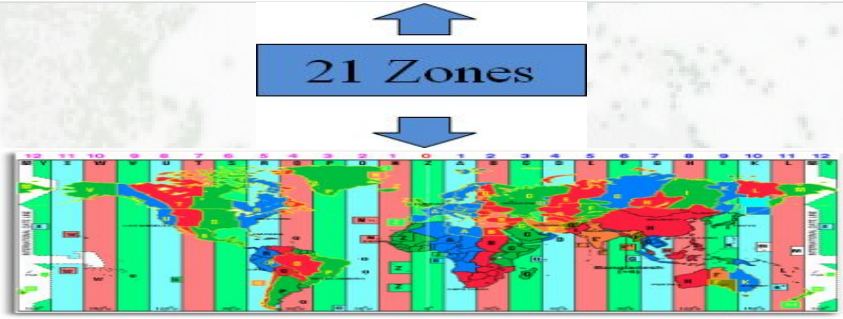


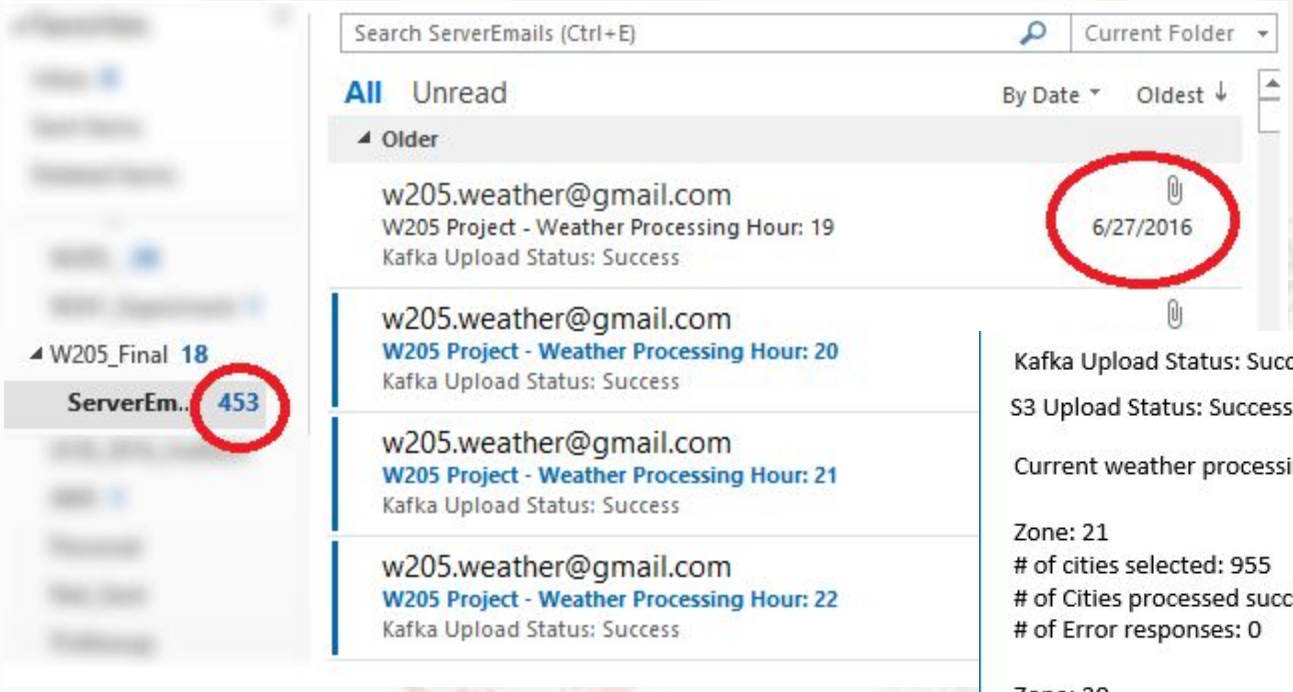
Implementation Details

Data Acquisition









Kafka Upload Status: Success

S3 Upload Status: Success

Current weather processing status by Zone:

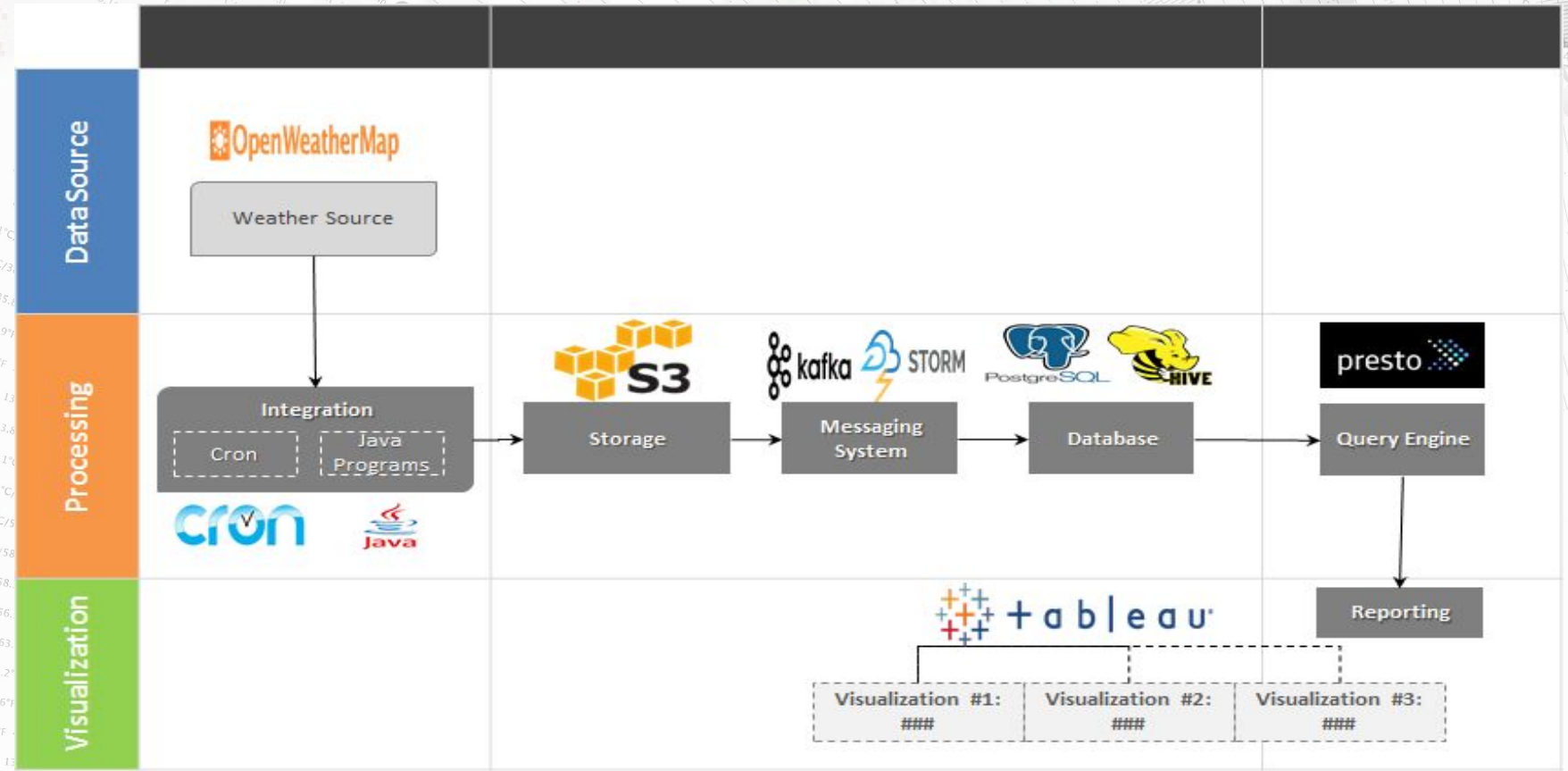
Zone: 21
# of cities selected: 955
# of Cities processed successfully: 955
# of Error responses: 0

Zone: 20
# of cities selected: 1484
# of Cities processed successfully: 1484 # of Error responses: 0

Zone: 19
# of cities selected: 4399
# of Cities processed successfully: 4346 # of Error responses: 53

Zone: 18
# of cities selected: 7321

Data Flow


Data Source
Processing
Visualization
OpenWeatherMap
Weather Source
Integration
Cron
Java Programs
cron
Java
S3
Storage
kafka
STORM
Messaging System
PostgreSQL
HIVE
Database
presto
Query Engine
Reporting
tableau
Visualization #1: ###
Visualization #2: ###
Visualization #3: ###

Data Velocity and Size

Data Velocity

Daily : 630,000 pulls
Per Minute : 440 pulls
Per Second : 6 pulls

Data Volume

Daily : 0.7GB
12 Weeks : ~60GB

Data Variety

Analysis

Analyses

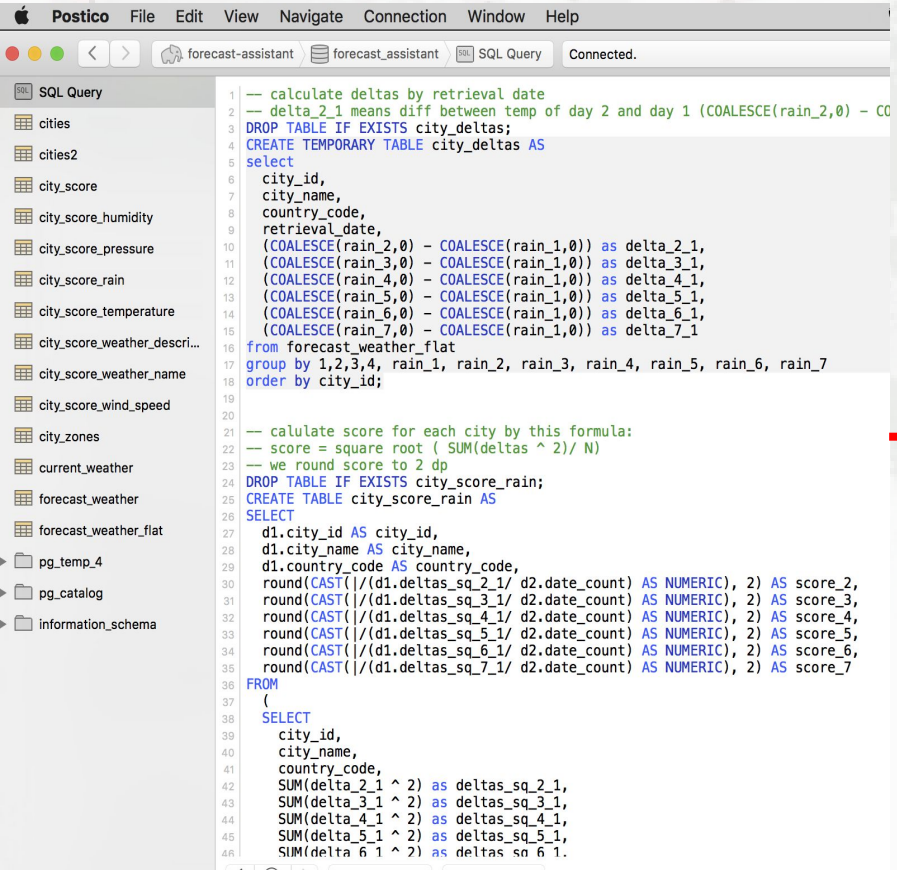
- Table views:
  - city_score_temperature, city_score_weather_description, city_score_weather_name, city_score_rain, city_score_pressure, city_score_humidity, city_score_wind_speed
- How to we score accuracy for each city?
  - For ‘numeric’ forecasts (temperature, humidity, rain, wind speed), score is the ‘variance’ of the forecasts from the actual data

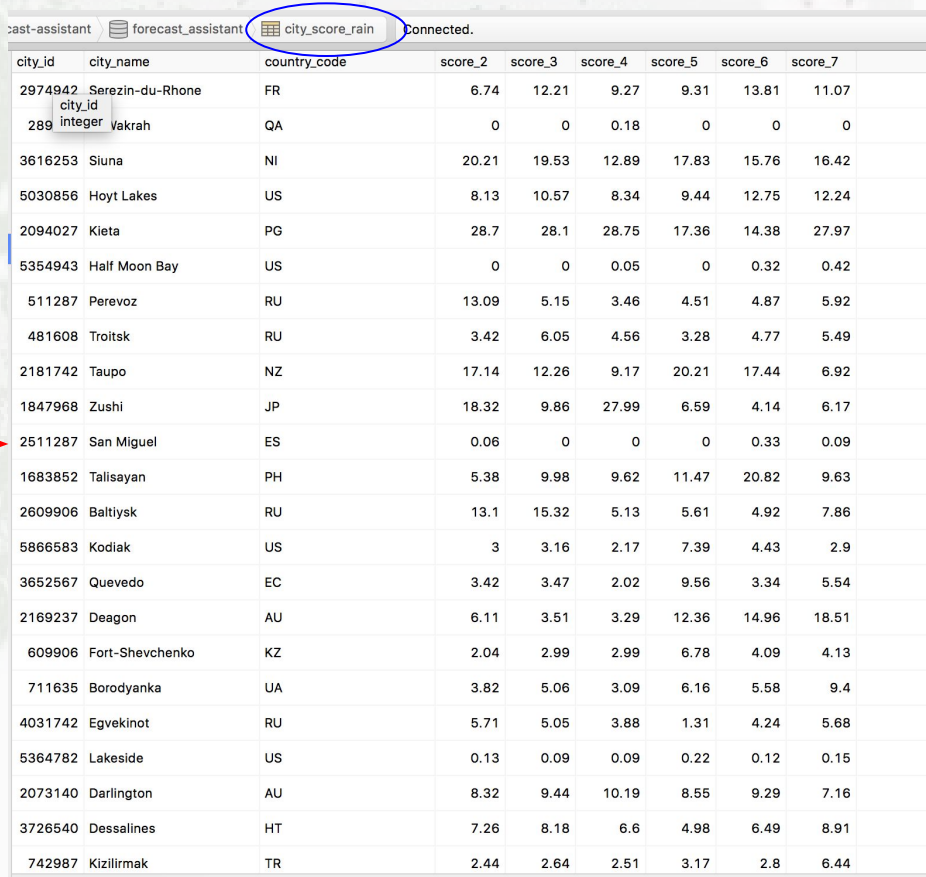
$$\text{Score} = \text{sqrt}((\text{sum}(\text{deltas} \wedge 2)/ N))$$

  - For ‘string’ forecasts (weather name, weather description), score is simply the % of times forecast is correct

$$\text{Score} = \text{accurate forecasts}/ \text{forecasts}\ (\%)$$

Table views: city_score_rain





city_id	city_name	country_code	score_2	score_3	score_4	score_5	score_6	score_7
2974942	Serezin-du-Rhone	FR	6.74	12.21	9.27	9.31	13.81	11.07
289	Vakrah	QA	0	0	0.18	0	0	0
3616253	Siuna	NI	20.21	19.53	12.89	17.83	15.76	16.42
5030856	Hoyt Lakes	US	8.13	10.57	8.34	9.44	12.75	12.24
2094027	Kieta	PG	28.7	28.1	28.75	17.36	14.38	27.97
5354943	Half Moon Bay	US	0	0	0.05	0	0.32	0.42
511287	Perevoz	RU	13.09	5.15	3.46	4.51	4.87	5.92
481608	Troitsk	RU	3.42	6.05	4.56	3.28	4.77	5.49
2181742	Taupo	NZ	17.14	12.26	9.17	20.21	17.44	6.92
1847968	Zushi	JP	18.32	9.86	27.99	6.59	4.14	6.17
2511287	San Miguel	ES	0.06	0	0	0	0.33	0.09
1683852	Talisayan	PH	5.38	9.98	9.62	11.47	20.82	9.63
2609906	Baltiysk	RU	13.1	15.32	5.13	5.61	4.92	7.86
5866583	Kodiak	US	3	3.16	2.17	7.39	4.43	2.9
3652567	Quevedo	EC	3.42	3.47	2.02	9.56	3.34	5.54
2169237	Deagon	AU	6.11	3.51	3.29	12.36	14.96	18.51
609906	Fort-Shevchenko	KZ	2.04	2.99	2.99	6.78	4.09	4.13
711635	Borodyanka	UA	3.82	5.06	3.09	6.16	5.58	9.4
4031742	Egvekinot	RU	5.71	5.05	3.88	1.31	4.24	5.68
5364782	Lakeside	US	0.13	0.09	0.09	0.22	0.12	0.15
2073140	Darlington	AU	8.32	9.44	10.19	8.55	9.29	7.16
3726540	Dessalines	HT	7.26	8.18	6.6	4.98	6.49	8.91
742987	Kizilirmak	TR	2.44	2.64	2.51	3.17	2.8	6.44

ast-assistant › forecast_assistant › city_score_temperature Connected.

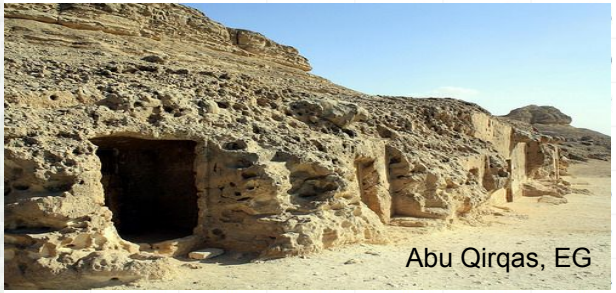
city_id	city_name	country_code	score_2	score_3	score_4	score_5	score_6	score_7
2094027	Kieta	PG	0.89	1.19	1.16	1.34	1.03	1.33
4031742	Egvekinot	RU	1.82	2.03	3.05	2.51	4.03	4.24
2181742	Taupo	NZ	4.86	4.1	5.93	6.49	6.37	5.26
711635	Borodyanka	UA	4.18	4.79	5.57	6.56	7.45	7.33
4883207	Alsip	US	3.39	4.82	5.11	5.27	5.03	4.52
742987	Kizilirmak	TR	4.02	4.69	5.25	5.53	6.04	5.65
1847968	Zushi	JP	2.65	2.84	2.74	2.83	2.59	2.3
3726540	Dessalines	HT	5.85	6.66	8.76	10.32	10.97	10.81
5030856	Hoyt Lakes	US	4.59	4.48	4.28	5.47	4.89	6.29

1

ast-assistant › forecast_assistant › city_score_weather_description

city_id	score_2	score_3	score_4	score_5	score_6	score_7
289915	0.94	1	0.94	0.88	0.88	0.94
362277	1	1	1	1	0.94	0.94

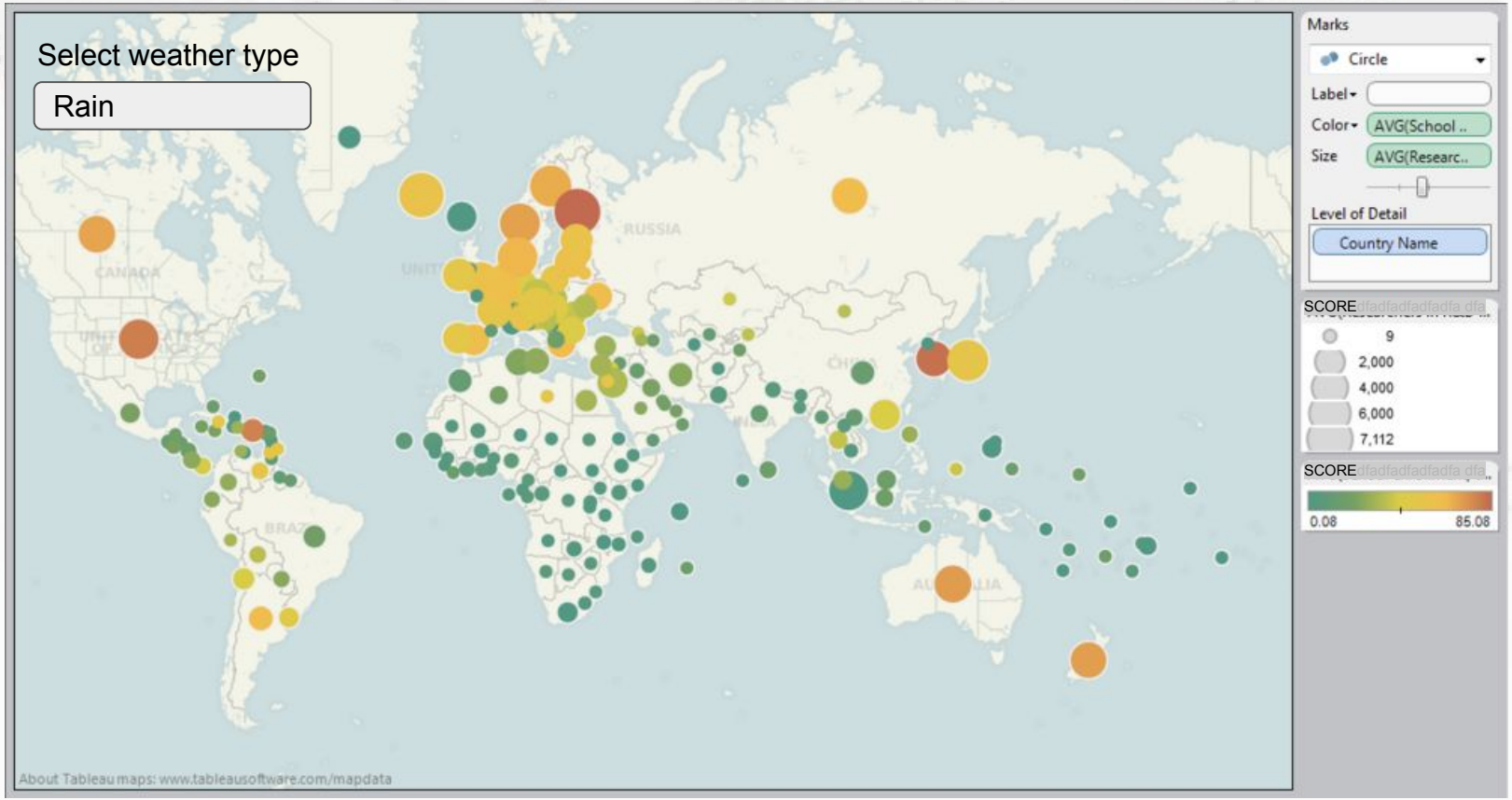
2



2073140	0.29	0.12	0.18	0.41	0.24	0.24
2094027	0.18	0.18	0.18	0.18	0.06	0.18
2169237	0.4	0.4	0.4	0.33	0.33	0.2
2181742	0.2	0.2	0.27	0.27	0.07	0.2
2511287	0.55	0.45	0.45	0.36	0.45	0.64
2609906	0.07	0.36	0.14	0.21	0.36	0.21
2974942	0	0.13	0.2	0.27	0.07	0.2

3

Map the world with weather scores



Project Conclusion

- Quality of weather forecast varies based on the location and the type of forecast
- We demonstrated that this is valuable information to be able to attach a numeric value to how much we should trust a particular weather forecast
- Hopefully, this would bring value to tourists, construction workers, farmers, campers and anyone who watches the weather to plan their daily activities

Postmortem

What went well?

- Team meetings every week since May.
- Setup Slack.com channel to improve team communications.
- Project played to team members individual strengths.

What could have gone better?

- More time to experiment with technology and to develop greater competencies.
- Longer semester to tweak production system and to play with analysis.

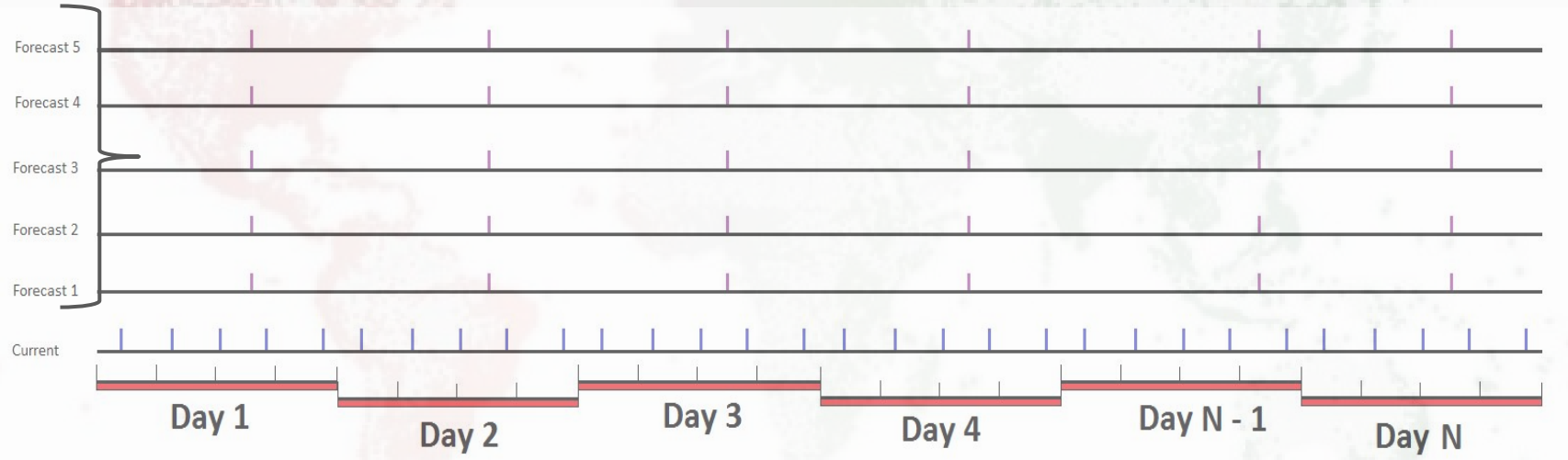
Future Work

- Collect forecast data at 3-hour intervals so that we could have more accurate and granular weather evaluation
- Include data for a variety of sources to be able to have more comprehensive comparison of the quality of different forecasters
- Collect data for a longer period of time to be able to evaluate accuracies at different times of year
- Build:
  - REST API
  - Web Page
  - Javascript library

Q&A

Q &A

Q&A


Forecast 5
Forecast 4
Forecast 3
Forecast 2
Forecast 1
Current
Day 1
Day 2
Day 3
Day 4
Day N - 1
Day N