
National Aeronautics and Space Administration
1
SHERLOCK
Sherlock Data Warehouse
NASA Ames, USRA/Crown/FRA, SGT/ATAC
June 2018
SHERLOCK

Objectives

- Introduce Sherlock to those who are new to it
- Show you Sherlock 2.0: new features to get you to the to core of your analysis
- Gather your ideas on future enhancements

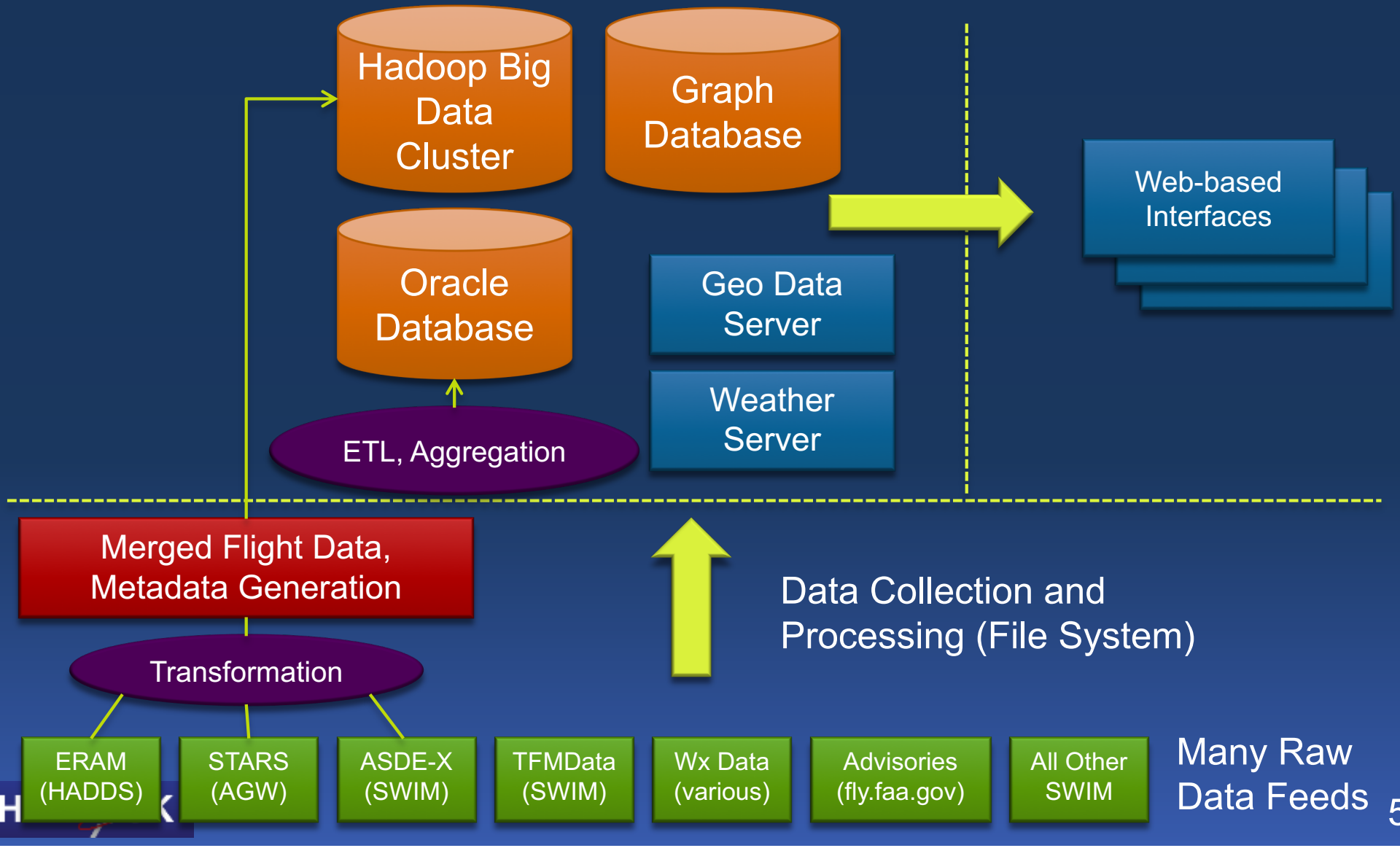
Outline

- Sherlock Overview
- Current Web Interface and Data Sources
- ATAC End-to-End Flight Data
- ATAC Reports and Aggregated Reports
- Big Data System
- Example Uses of Sherlock for Analysis
- Semantic Graph Database
- Resources/Backup

Sherlock Overview

- NASA has access to many ATM flight, weather, and traffic flow-related data sources
- Real data drives all of NASA's ATM R&D
- Sherlock is a platform for reliable ATM data collection, archiving, processing, query, and delivery
- Sherlock is a platform for big data analytics, including data mining and machine learning
- Sherlock also has a semantic data store aimed at enabling more complex data integration

Simplified Sherlock Architecture



Major Sherlock Functions

- Search and Download of raw data – e.g., access to data collected from original sources and written to file system. Ex: FAA SWIM, NOAA Weather
- Browse, Query Download parsed sources such as METAR, or derived data, such as WITI
- Big Data server supporting analytics of track and weather data on Hadoop platform
- Graph database supporting complex queries across heterogeneous data sources
- Geospatial server to query and visualize airspace elements (routes, fixes, boundaries, etc.)
- Weather server to query and visualize weather

Recent Enhancements – Sherlock 2.0

- Addition of all available FAA SWIM data feeds
  - SWIM is FAA's new way of distributing NAS data
  - All feeds are in XML format, but no two feeds use same schema. (Center schema not same as TRACON)
- Addition of ATAC processed flight data
  - All data in same, consistent format (csv records)
  - Data stored by facility (TRACON/Surface facilities paired)
  - Data also merged into end-to-end flight records over entire US National Airspace
- Addition of ATAC reports
  - Daily reports based on flight data, FAA advisories, etc.
- Advances in Big Data system (Cloudera Hadoop)
  - ATAC data being loaded into HDFS cluster daily
  - SMARTNAS API to big data
  - Used for NRA research

FAA SWIM Data Sources on Sherlock

- Flight Data:
  - STDDS/ASDE-X: Surface data
  - TAIS: TRACON data, including VFR flights
  - SFDPS: Center data from ERAM
  - TFMData: NAS-wide flight data, flow constraints
  - TBFM: Operational metering data
- Airport Data:
  - ADPS: Airport Data Service, Runway Visual Range info
  - NOTAM: Notices to Airmen
- Weather Data:
  - ITWS: terminal convective weather
- SWIM Schema Info

Other Raw Data on Sherlock

- Weather:
  - CIWS convective forecasts
  - METAR airport current surface weather conditions
  - NOAA Rapid Refresh (RR) forecasts
- Flight Data:
  - CTAS text-based format, in Center/TRACON pairs
- Stored but no longer updated (obsolete):
  - NOAA RUC forecasts
  - ASDI flight data

Processed Data on Sherlock

- OPSNET Stats
- ATCSCC Strategic Advisories
- METAR Airport Weather Reports
- TAF Weather Forecasts
- PIREP Pilot Reports
- WITI Weather Impact Analysis by day, Center, and Sector (computed by Sherlock)
- CCFP Simplified Weather Polygons

New Processed Data from ATAC

- Collected for 76 FAA facilities. Per-facility data available next day. Merged USA data available within 7 days. Consistent format
- Flight Data
  - IFF: Flight Plan and Track data
  - EV: Flight Event Data
  - RD: Flight Summary Data
- Reports (to be discussed later)
  - Go Arounds
  - Turn to Final
  - Runway Usage
  - Taxi Time
  - Field 10 Reroutes
  - Center Instantaneous Counts
  - Sector Statistics
  - Sector Activity
  - Best Flight Plan

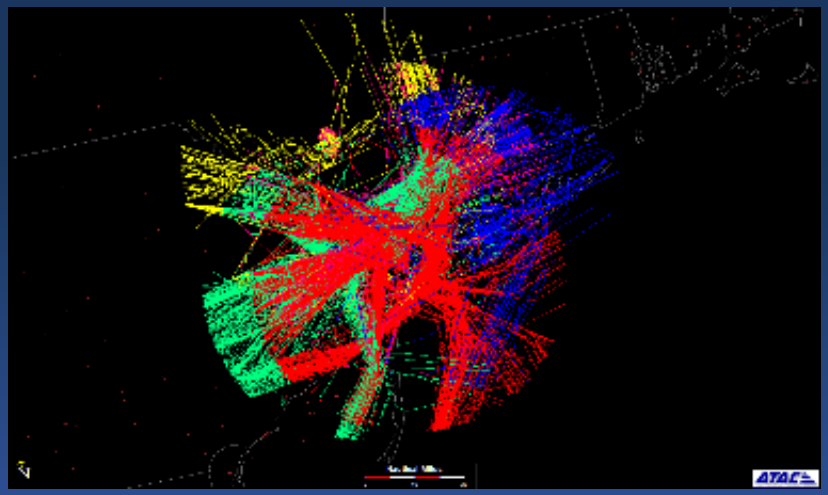
ATAC Flight Data

- IFF: Integrated Flight Format
- Includes all source data plus derived fields
- Record types include Flight Summary (2), Track Point (3), Flight Plan (4):
  - Summary: Time, Key, Beacon, Source, AC ID, AC Type, Orig, Dest, Ops Type
  - Track: Time, Key, Beacon, Source, AC ID, Lat, Long, Alt, Accuracy, Ground Speed, Course, Rate of Climb, Facility, Mode S, etc.
  - Plan: Time, Key, Beacon, Source, AC ID, AC Type, Orig, Dest, Altitude info, Route, ETA, Flight Cat, Perf Cat, Ops Type, Equipage, Coord Time, etc.
- Complete IFF Specification

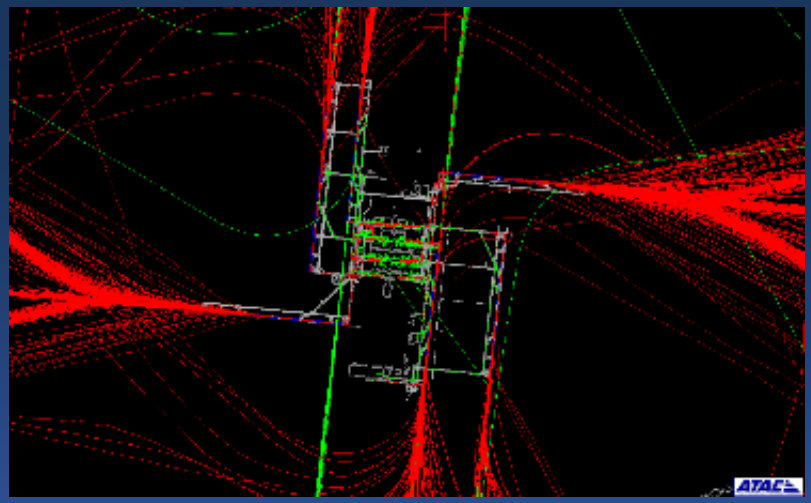
Flight Data Availability

STARS –available from 9/1/2015

ERAM –available from 10/1/2015

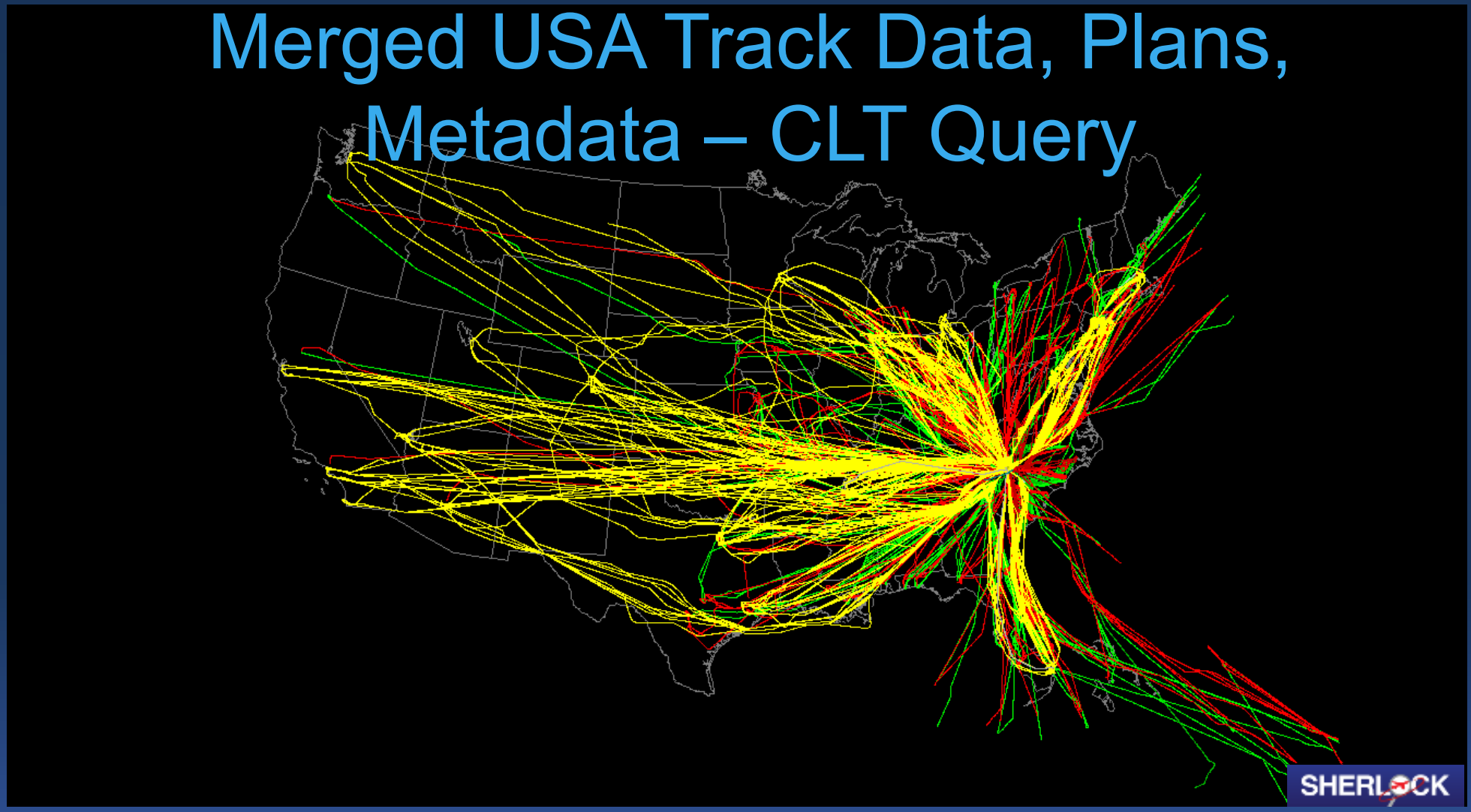


ARTS –available from 11/1/2015



ASDE-X –available from 1/16/2016

Merged USA Track Data, Plans, Metadata – CLT Query



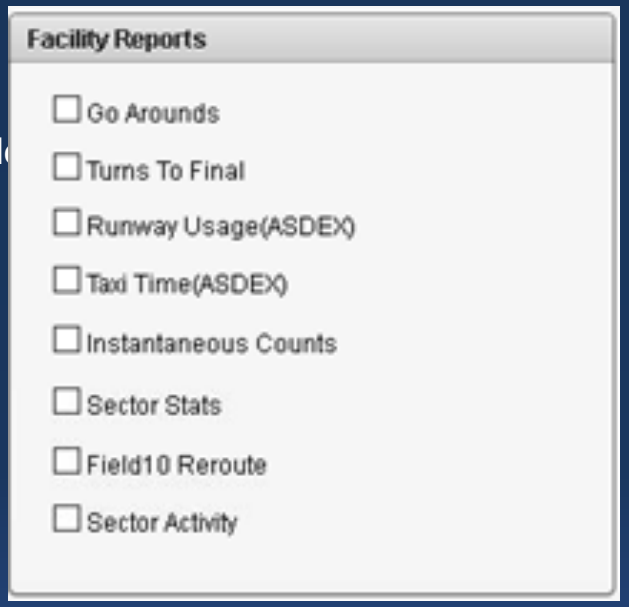
End to End (N2N) – available from 1/16/2016

ATAC Summary Data

- RD: Reduced Data summary, one record per flight
- In USA merged, departure and arrival fields are populated
- Fields: Key, Track Start, Track End, Duration, AC ID, AC Type, Beacon, Ops Type, Airline, Carrier Type, Origin, Destination, Takeoff Runway, Landing Runwy, Top of Climb, Top of Descent, Takeoff Time, Landing Time, Route, Traversed: Centers, TRACONS, Sectors, and SUAs, etc.
- Complete RD Specification

ATAC Reports in Sherlock

- Performance Reports – available daily for individual facilities
- Terminal (STARS, ARTS)
  - Go around reports
    - Counts, runways, altitude, return time
  - Turn-to-Final
    - Overshoots, glideslope speed/altitude deviations, turn on angl
- Surface (ASDE-X, ASSC)
  - Runway Usage
    - Runway throughput, arrival/departure rates
  - Taxi – time
    - Taxi out, taxi in time
- En-route (ERAM)
  - Instantaneous Counts Reports (static and dynamic)
  - Sector Stats (static and dynamic)
  - Reroutes
  - Sector Activity
- NAS – wide
  - Best Flight Plan (Synthesized)
  - CCFP Sector Coverage
  - CWAM Sector Coverage



ATAC Aggregated Databases

- Aggregated Trend Databases – Updated Daily
  - En-route
    - Sector Activity
    - Reroutes
    - Sector Stats
    - Instantaneous Counts
  - Terminal
    - Go arounds
    - Turn to Final
  - Surface
    - Runway Usage
    - Taxi Time
  - NAS Wide
    - Best Flight Plan
- Data sets Aggregated Monthly
  - CCFP Sector Coverage
  - CWAM Sector Coverage

Summary: ATAC Processing

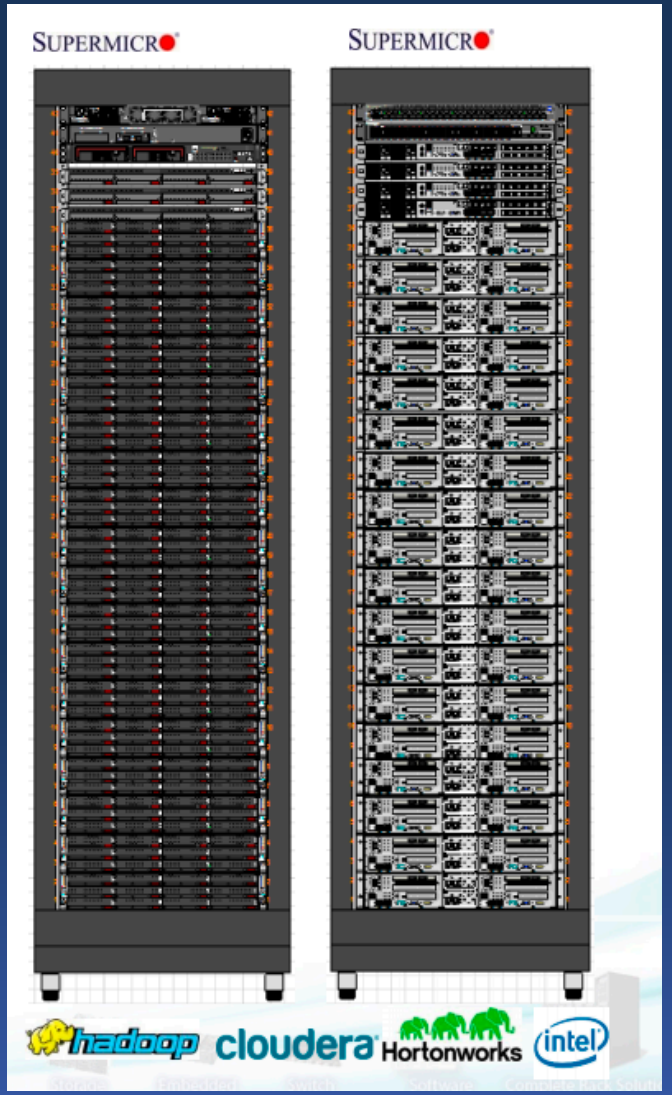
Number	Description
8	Dedicated Processing Machines
76	ATC Facilities collected/processed daily
39,000+	Analysis ready facility-days processed
116,000+	Number of performance reports generated to date
40-50 million	Number of track points in a 24-hour set of end-to-end data
infinite	Thanks to NASA and contractor IT/Lab Staff that made this possible

What is a Hadoop Big Data System?

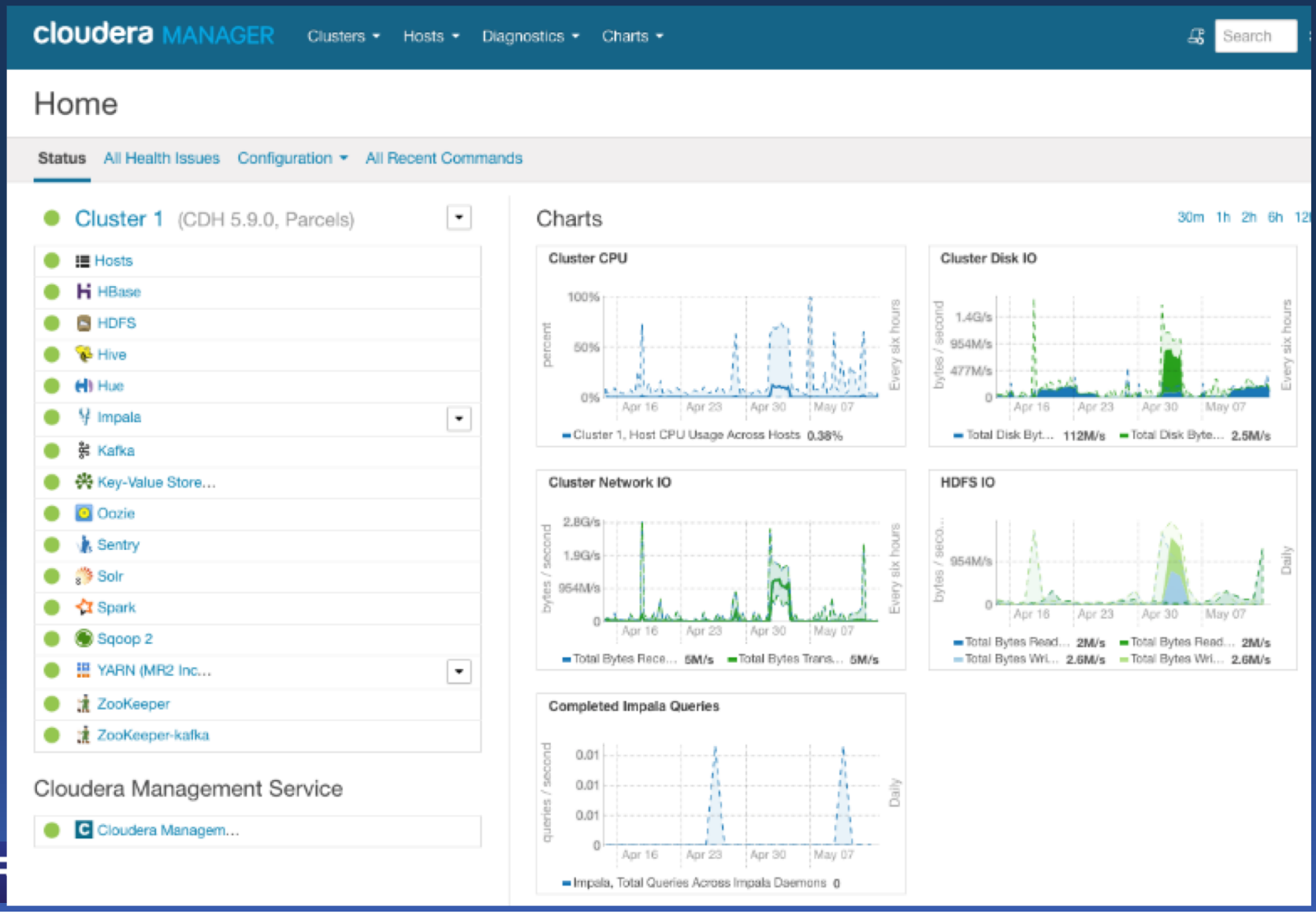
- Hadoop is an open-source software framework for storing data and running applications on clusters of commodity hardware. It provides massive storage for any kind of data, enormous processing power and the ability to handle virtually limitless concurrent tasks or jobs.
- Many resources to learn Hadoop. Recommend Coursera series from UC San Diego: https://www.coursera.org/specializations/big-data

Sherlock Big Data System

- SuperMicro Engineered System
- Cloudera Hadoop software
- 42U rack
- Total of 576 CPU Cores, 800 TB Storage
- 1 Management Node
- 3 Name Nodes (Dual 8 Core, 512 GB RAM each)
- 32 Data Nodes (Dual 8 Core, 256 GB RAM each)



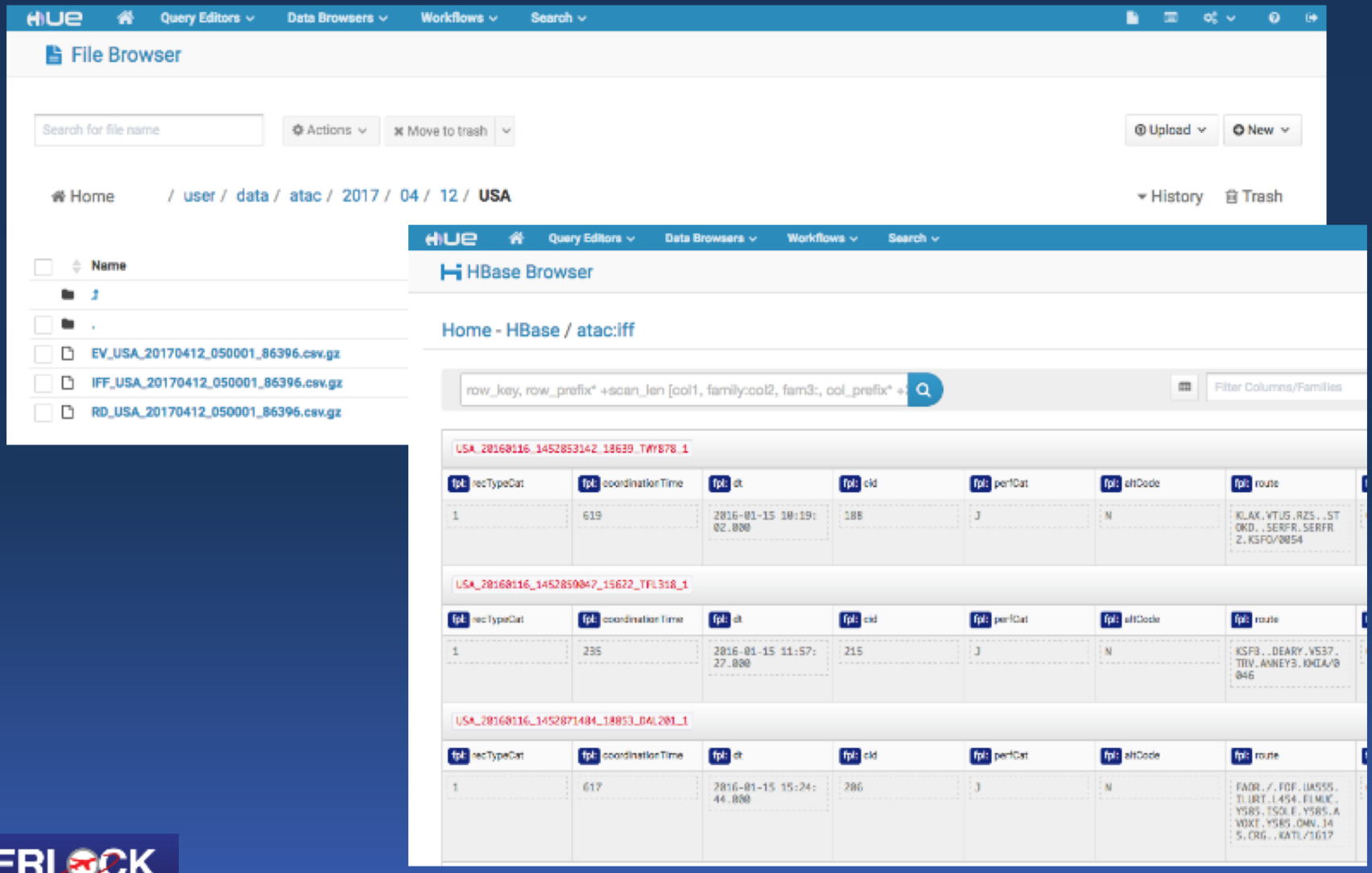
Sherlock Big Data Services



Big Data Status

- Cloudera stack installed and running; data imported daily
- User interface is here: http://sherlock.arc.nasa.gov:8889/home
- Send email to sysadmin@osprey if you would like to try it out
  - Requires NextGen ATM account;
- HDFS data: populated with facility and USA flight data (IFF, EV, RD)
- HBASE data: populated with USA flight data (IFF, EV, RD) and facility data for sectorization
- We are looking for use cases to know what else to store there

Big Data Quick Demo



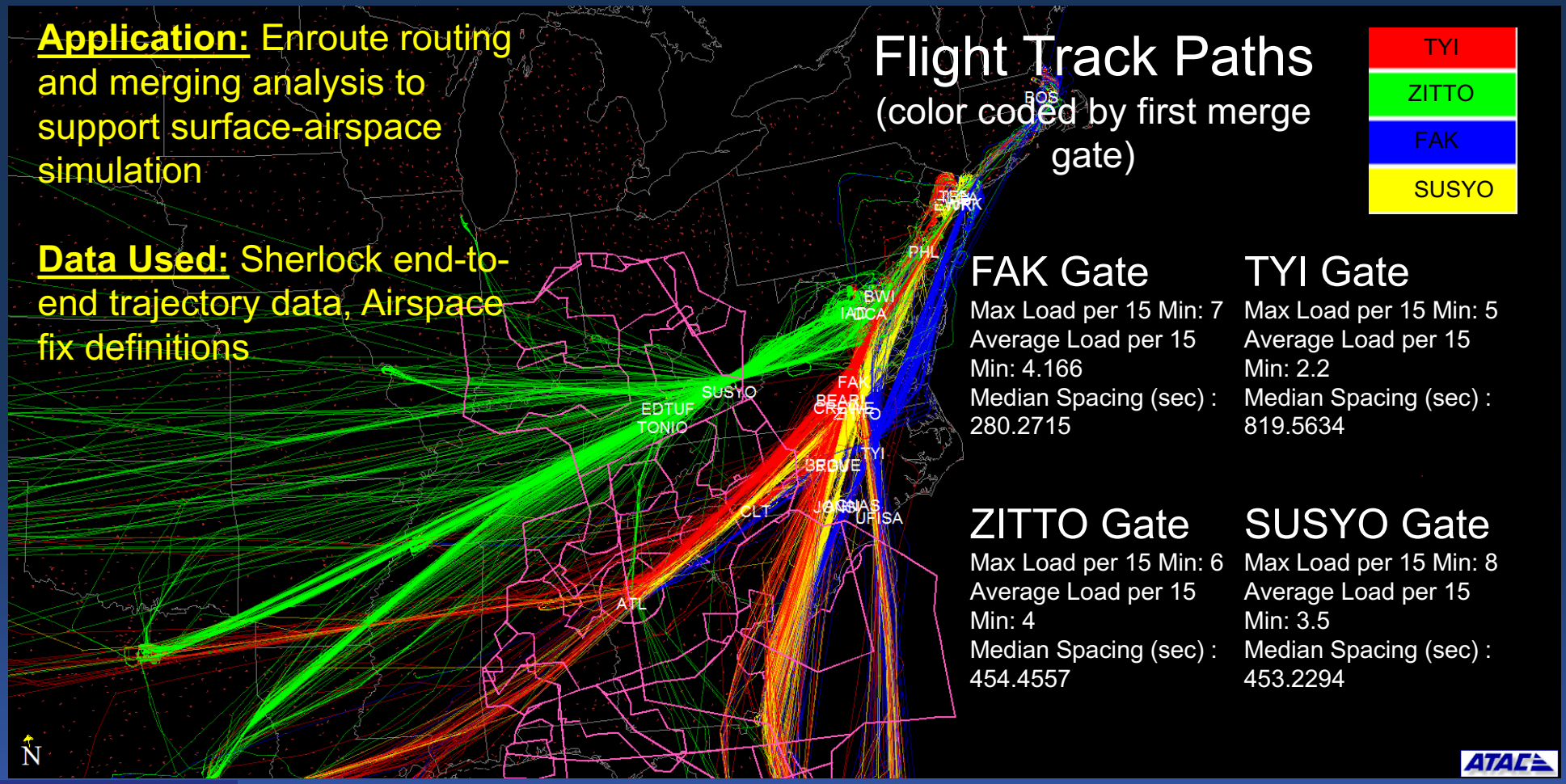
ATAC: Examples of Analysis Use

Name	Project	Sherlock Capability Used	Description
ATD-2 Benefits Assessment	ATD-2	End-to-end trajectories, surface flight data	Simulation validation, metrics dashboards
MFCR	ATD-3	End-to-end trajectories, CWAM, CIWS weather	Weather rerouting
Big Data Analytics for Aeronautics	SMARTNAS	Merged trajectories, metadata	Anomaly detection using Sherlock track data
Big Data Analytics for RSSA	ATAC SBIR	End-to-end trajectories, Flight summary, flight events, CWAM polygons, Sherlock big data system	Complex geospatial comparisons, complex search, real time analytics

Surface-Airspace Simulation For ATD-2 Benefits Analysis

Application: Enroute routing and merging analysis to support surface-airspace simulation

Data Used: Sherlock end-to-end trajectory data, Airspace fix definitions



FAK Gate

Max Load per 15 Min: 7
Average Load per 15 Min: 4.166
Median Spacing (sec) : 280.2715

TYI Gate

Max Load per 15 Min: 5
Average Load per 15 Min: 2.2
Median Spacing (sec) : 819.5634

ZITTO Gate

Max Load per 15 Min: 6
Average Load per 15 Min: 4
Median Spacing (sec) : 454.4557

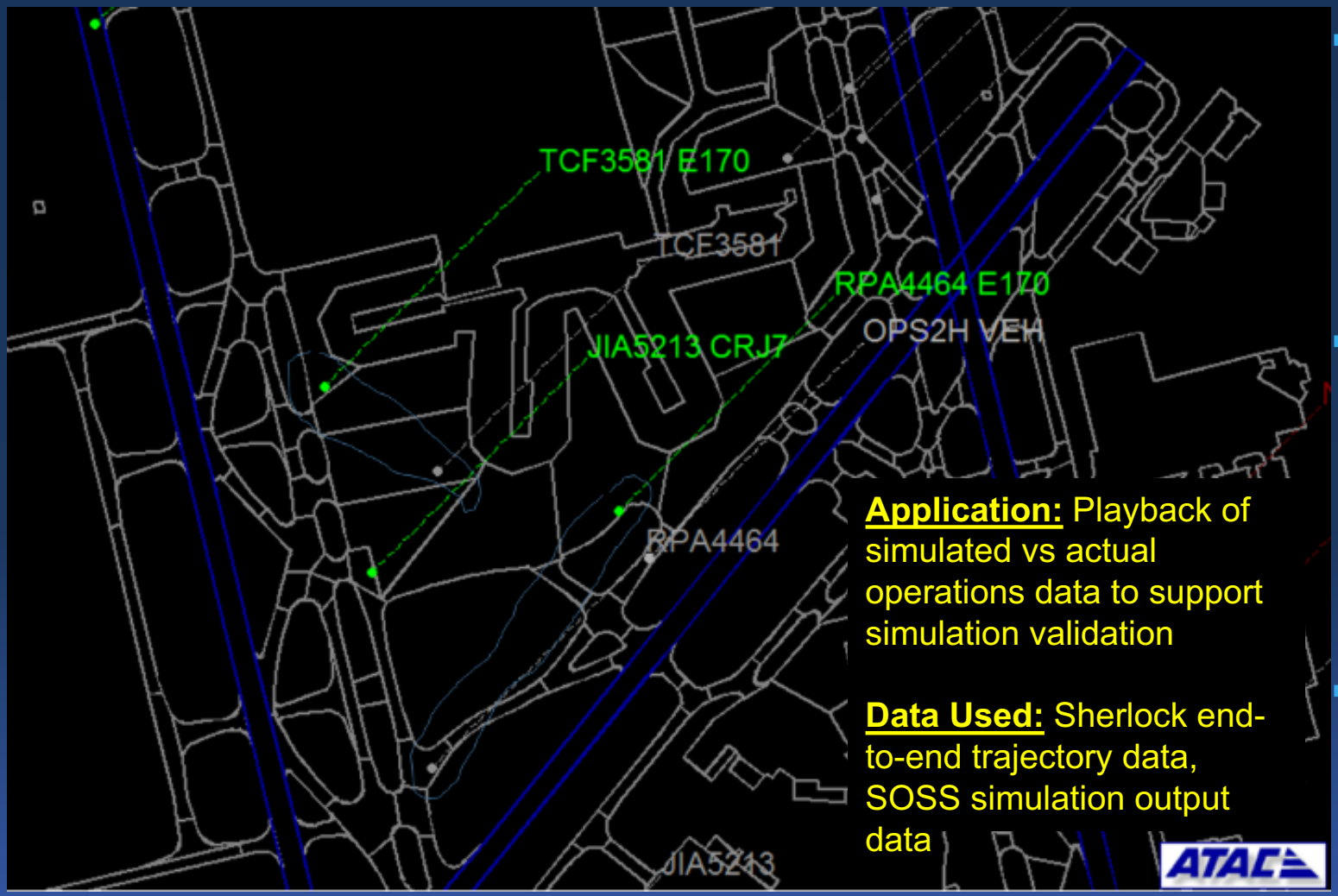
SUSYO Gate

Max Load per 15 Min: 8
Average Load per 15 Min: 3.5
Median Spacing (sec) : 453.2294

ATAC

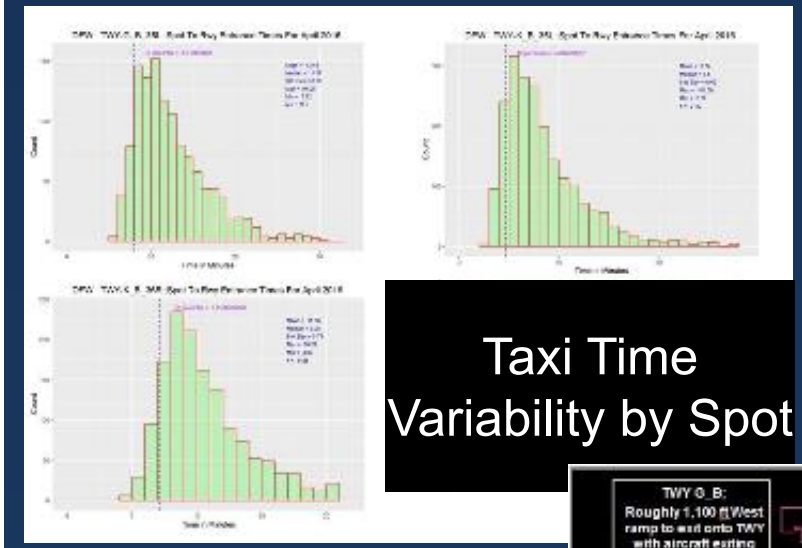
SHERLOCK

Fast Paced Simulation Validation and Post Analysis Process



- Convert simulation output data into ATAC/ Sherlock ASDE-X track data formats
- Develop new/ reuse existing Sherlock reports for creating a validation and post-analysis dashboard
- Playback capability to compare simulated flights with real flights

Historical Departure Operations Characterization

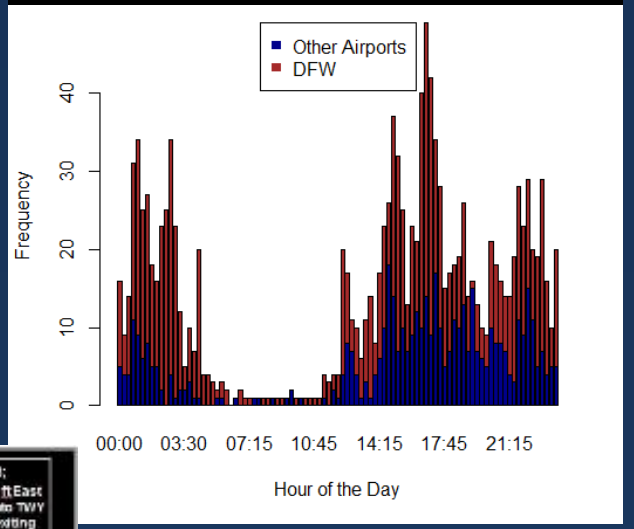


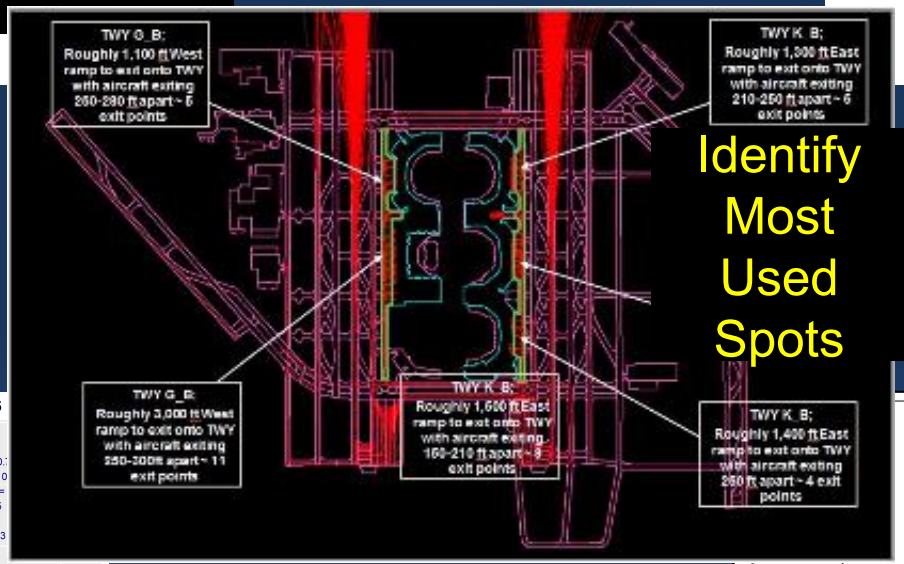
Taxi Time Variability by Spot

Application: Characterization of departure operations at study metroplex sites

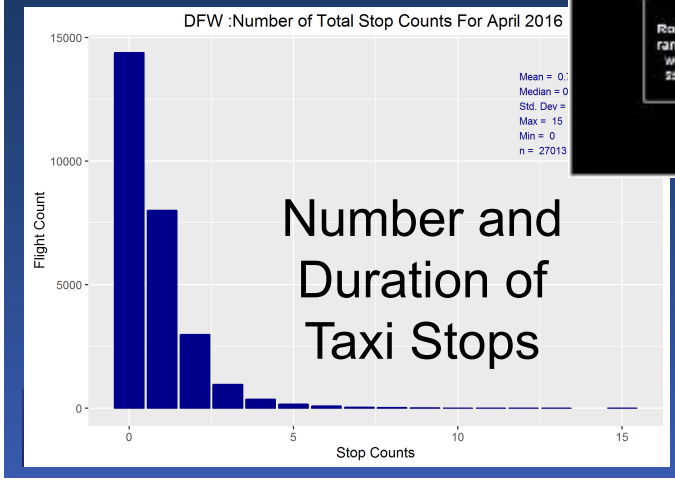
Data Used: Sherlock end-to-end trajectory data, airport layout definitions, Airspace fix definitions

Departure Fix Sharing

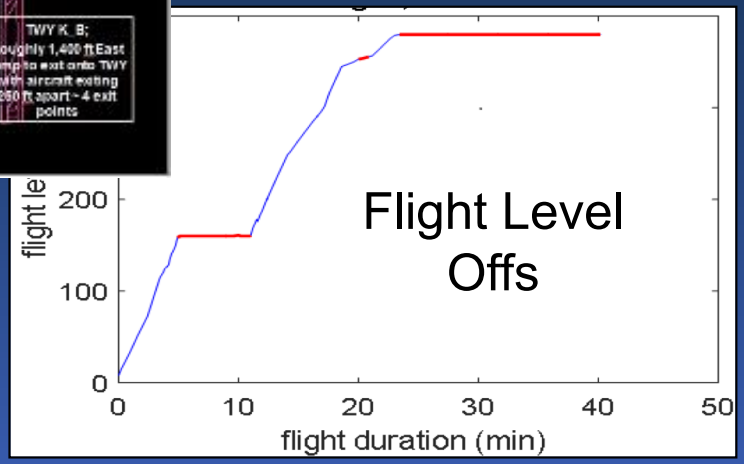




Identify Most Used Spots

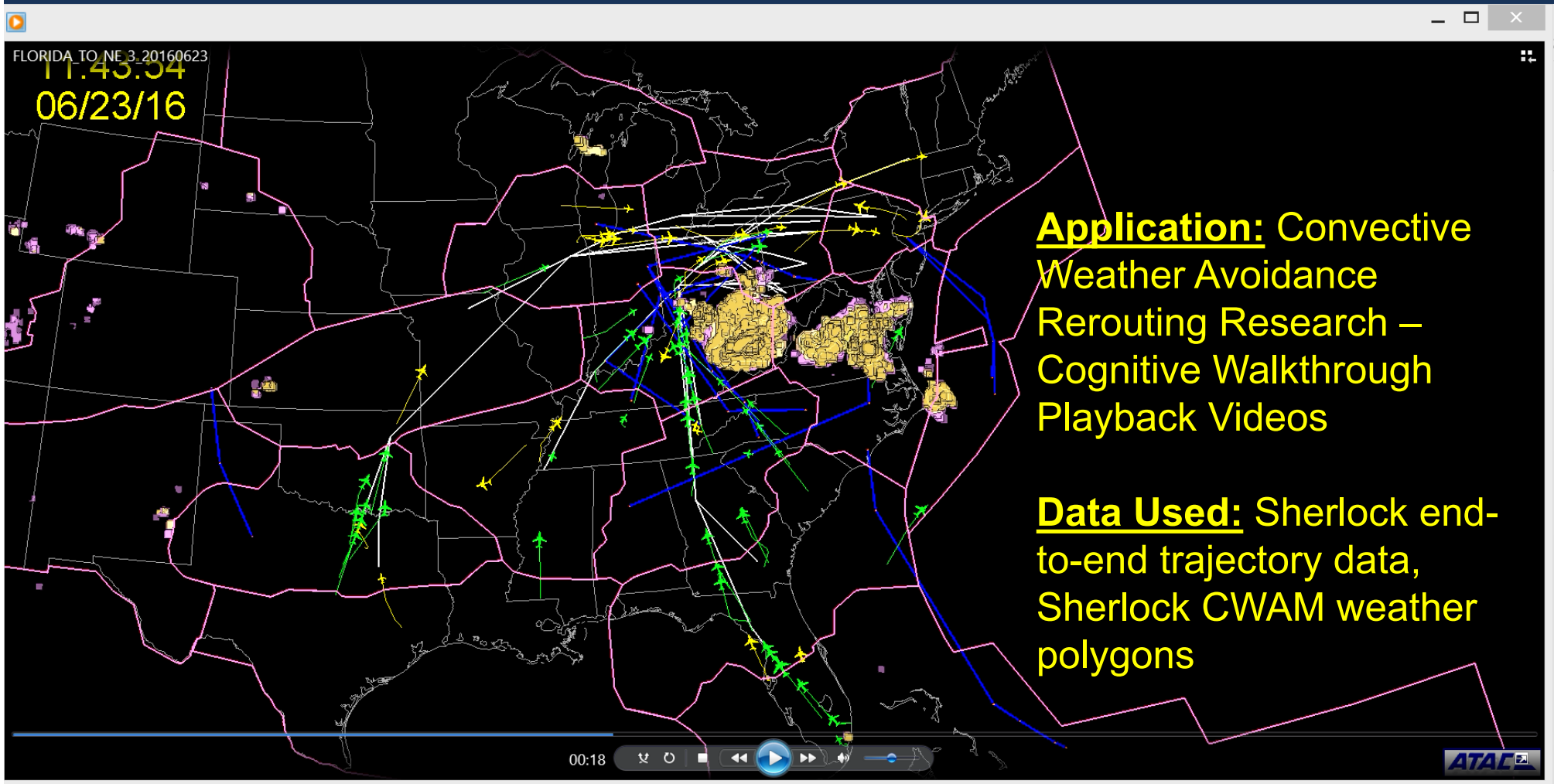


Number and Duration of Taxi Stops



Flight Level Offs

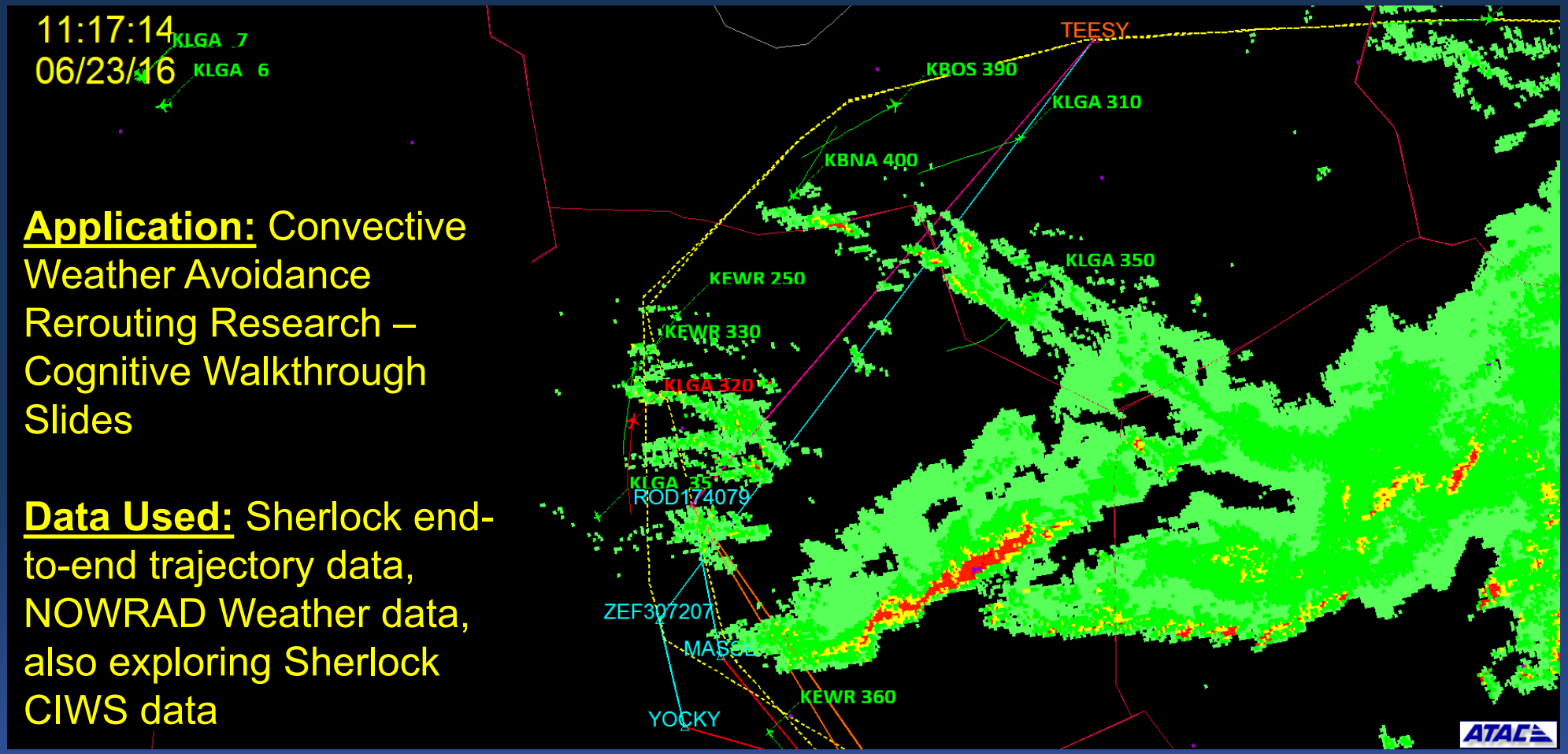
Convective Weather Rerouting Scenario Analysis



Application: Convective Weather Avoidance Rerouting Research – Cognitive Walkthrough Playback Videos

Data Used: Sherlock end-to-end trajectory data, Sherlock CWAM weather polygons

Convective Weather Rerouting Scenario Analysis



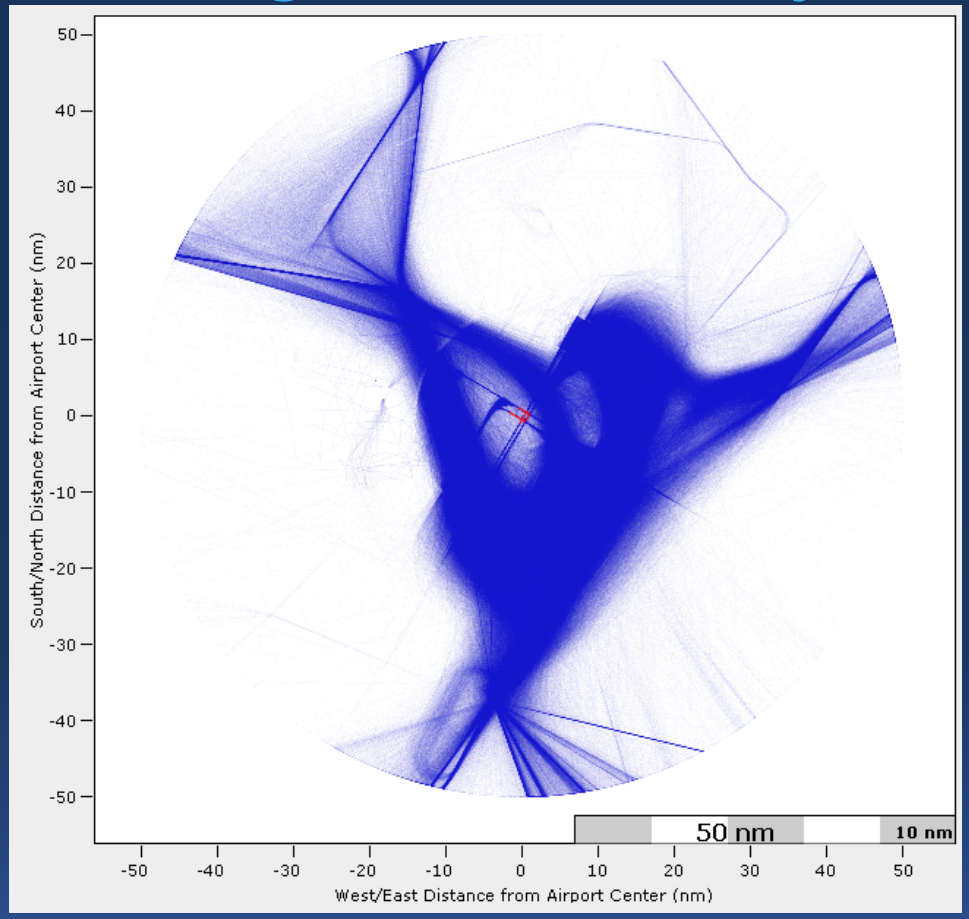
Application: Convective Weather Avoidance Rerouting Research – Cognitive Walkthrough Slides

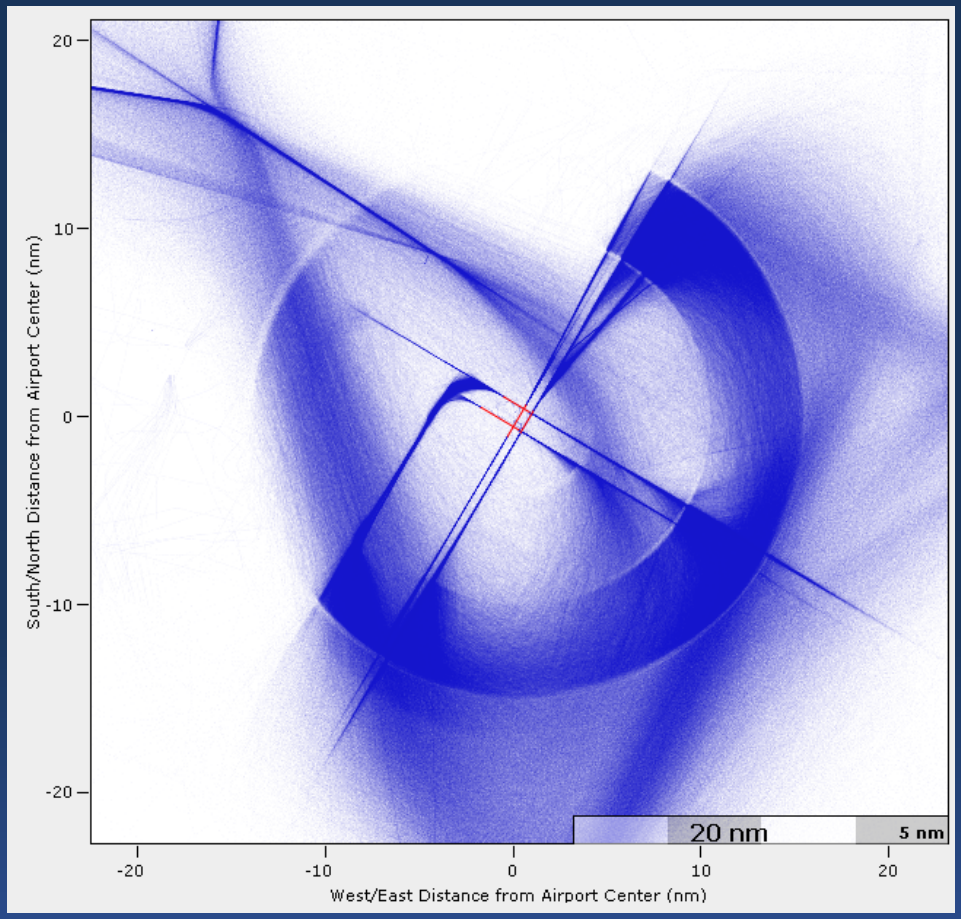
Data Used: Sherlock end-to-end trajectory data, NOWRAD Weather data, also exploring Sherlock CIWS data

Big Data Applications – ATAC Research

- Geospatial Computations on a large scale
  - Comparing CWAM polygons to end-to-end trajectory data
- Complex Search
  - Using SparkSQL, Apache Hive, MongoDB and other technologies to query years of operational data for specific NAS occurrences
- Real time analytics for RSSA use case
  - Detecting potential CWAM polygon proximity for NAS traffic in real time

Big Data Analytics for Aeronautics

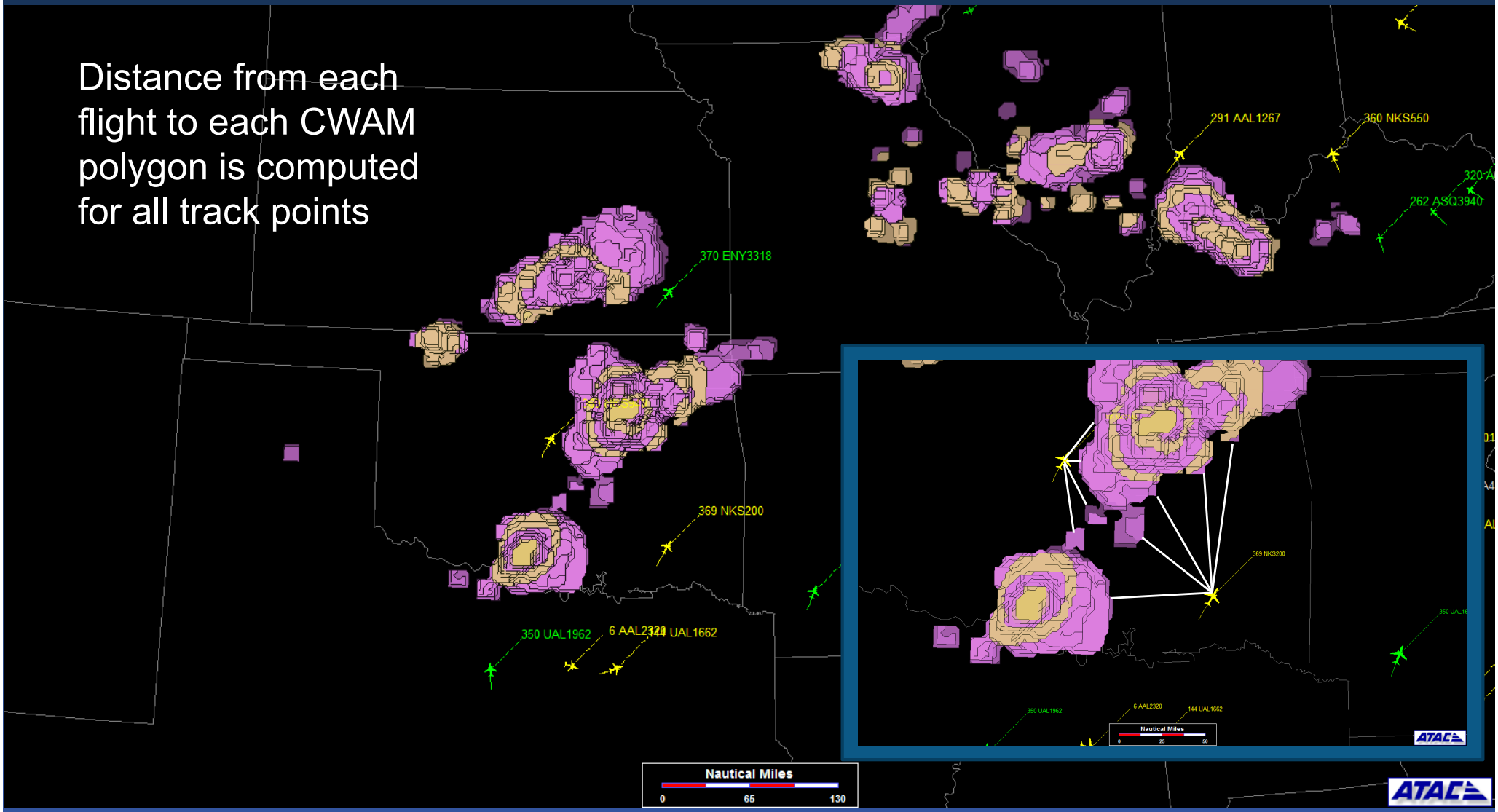




- Visualization of all ~40 million reported track points
- Frequent approach paths visible

Large Scale Geospatial Computation

Distance from each flight to each CWAM polygon is computed for all track points



ATAC Next Steps

- Underway
  - Historical Data Processing – En-route, Terminal data sets back to 1/1/2014
  - Additional Wx coverage and sector metrics for TBO project
  - Adding aggregated data to Oracle
  - Continued Big Data applications
- Wish List
  - VFR Flight Processing
  - Real-time data merging
  - Additional performance reports
  - Other?

Semantic Graph Database

- Dr. Rich Keller, Code TI

Wouldn't it be nice…

- To be able to query across *all* of the data sources in Sherlock to answer questions that cut across multiple kinds of data? (Not just the flight data…)
- To save much of the time and energy spent writing custom code to integrate data from multiple Sherlock datasets?

What are the challenges?

- Sherlock-housed datasets are very heterogeneous
  - data formats
  - field names
  - scientific units
  - spatial/temporal alignment
- Sherlock is a patchwork quilt
  - contains raw data files and structured Oracle & HBase tables
  - lacks field standardization across tables, making joins difficult
  - missing adaptation info. necessary to connect data tables

Result: Can only query within isolated ‘data islands’; can’t bridge across data sources without great effort

National Aeronautics and Space Administration

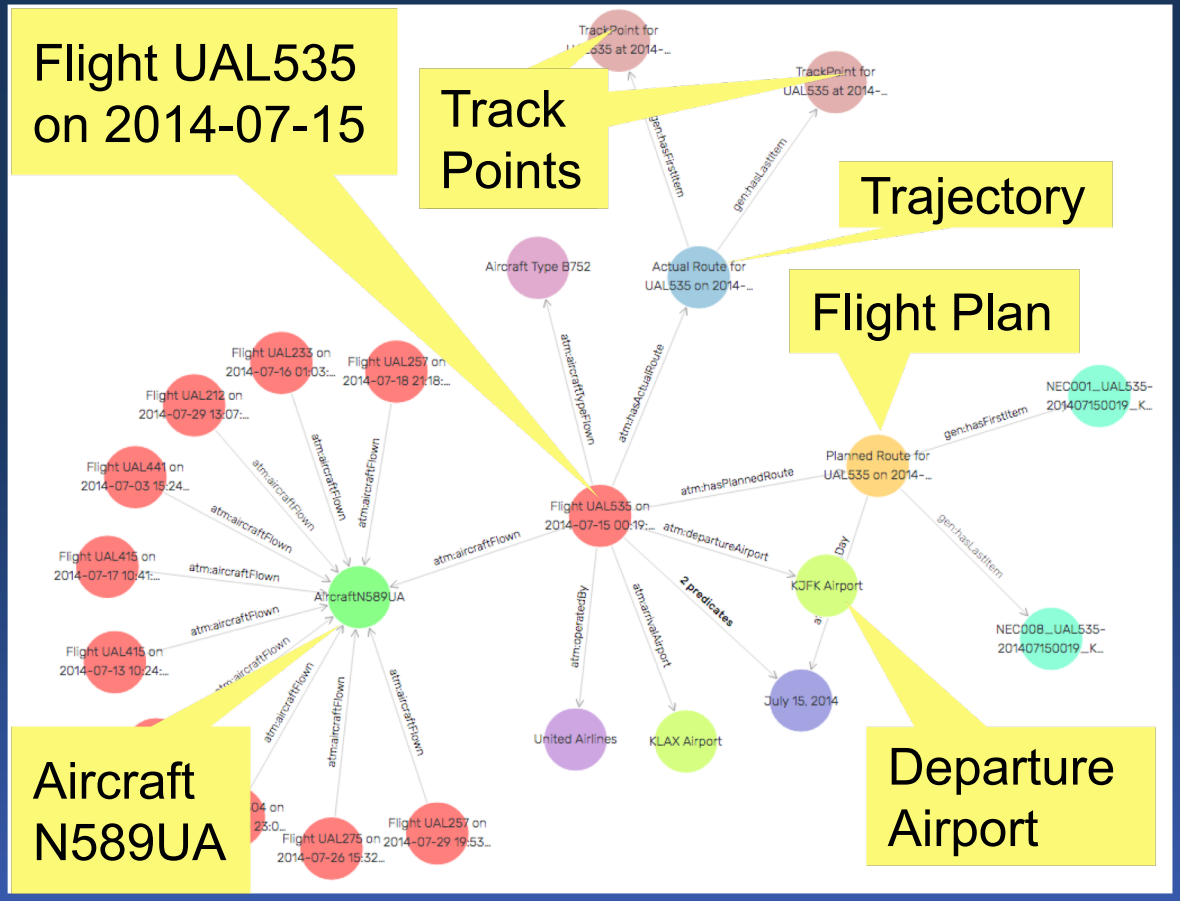
Ames

Discovery Innovations Solutions

Integrated Graph Database

- Highly-interconnected, network-structured info. store, where:
  - Nodes represent airspace entities and their properties
  - Links represent relationships

- Integrates all types of data within a single queryable structure
  - Flight track
  - Airport
  - Weather
  - Advisory
  - Aeronautical info.



SHERLOCK

Goal: Enable Cross-cutting Queries and data export capabilities

- Examples:
  - Find Delta flights using A319s departing ZNY airports in rain & heavy winds
  - Identify all sectors within which any A320 aircraft is currently operating in US airspace
  - During December 2014, locate all UAL flights that were rerouted due to weather, after departing ORD under freezing ground temperatures after ground delays of over 60 minutes.

Results to Date

- Evaluated feasibility of graph database approach 👍
- Designed an extensive graph data schema (an 'ontology')
- Acquired data from 8 heterogeneous data sources focusing on flight operations in the NY Metroplex (3 largest airports in NY area) during July 2014
- Loaded 100K flights into graph database (>38M nodes!)
- Designed and executed ~25 cross-cutting queries (with help of active AF and TI researchers) to support ongoing research
- Compared two different graph database products and measured query performance in a benchmarking exercise

To be continued… Stay tuned!

Resources

- Confluence Documentation: https://atmjira.arc.nasa.gov:9443/conf/display/ctas/Sherlock+Data+Warehouse+Home
  - We can export info above for external users
- Contacts:
  - Heather.Arneson@nasa.gov
  - Michael.E.LaScola@nasa.gov
  - Pallavi.Hegde@nasa.gov
  - Rich.Keller@nasa.gov
  - Jes@atac.com

Acknowledgements

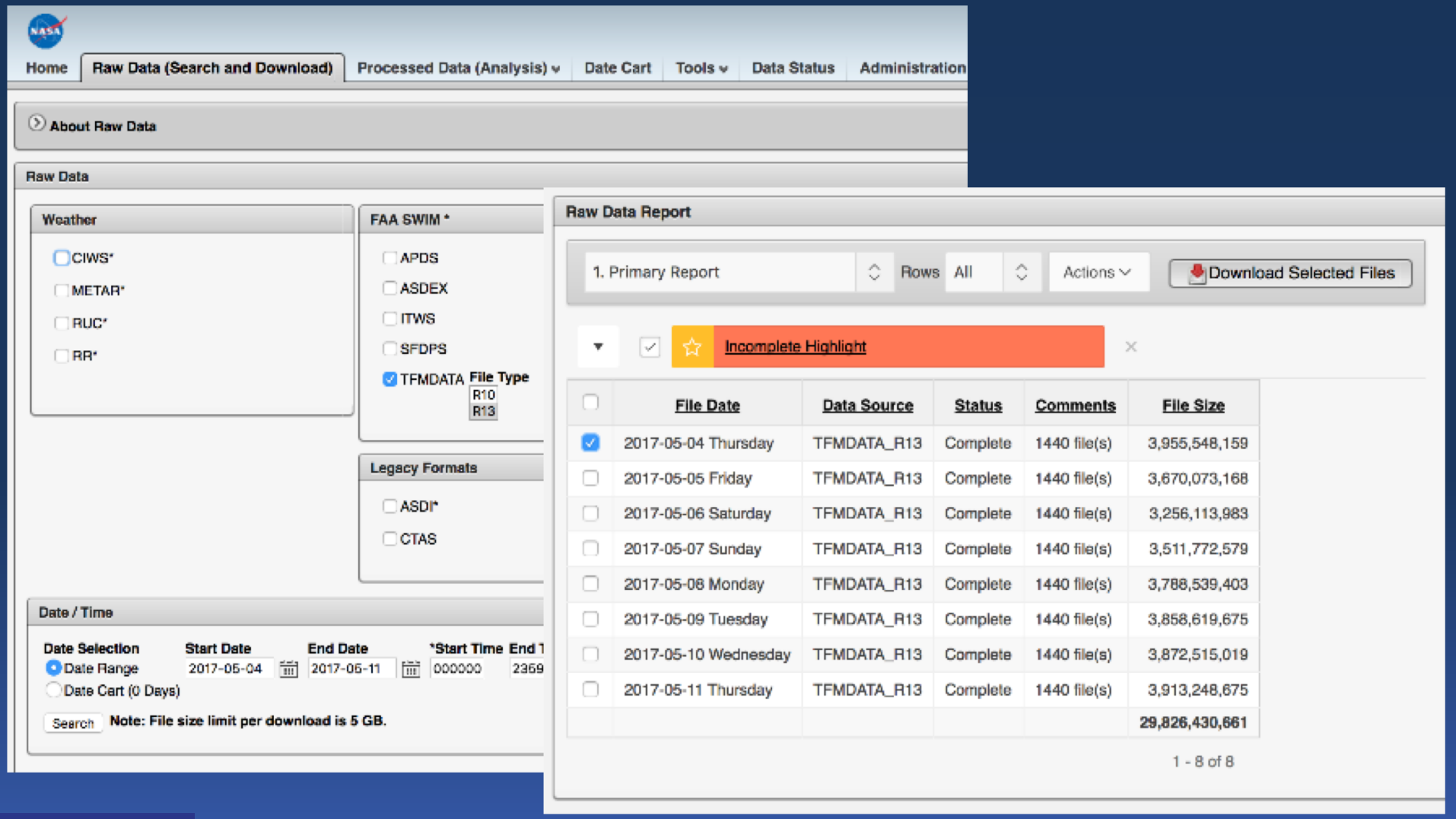
- Many people contribute to Sherlock!
- Developers: Michael La Scola, Pallavi Hegde, Shubha Ranjan (emeritus)
- Database & Big Data admin: Eric Wang, Dat Duong
- Data collection, archiving, monitoring: Pat O'Neal, Joe Cisek
- Windows, Linux admin: Matt Ma
- Graph database: Rich Keller, Mei Wei
- ATAC data: John Schade, Kennis Chan, Cindy Wong, Other Eric Wang, et. al.
- And thanks to Code AF for funding Sherlock, and to SmartNAS for supporting ATAC's work!

Backup Charts and Screen Shots

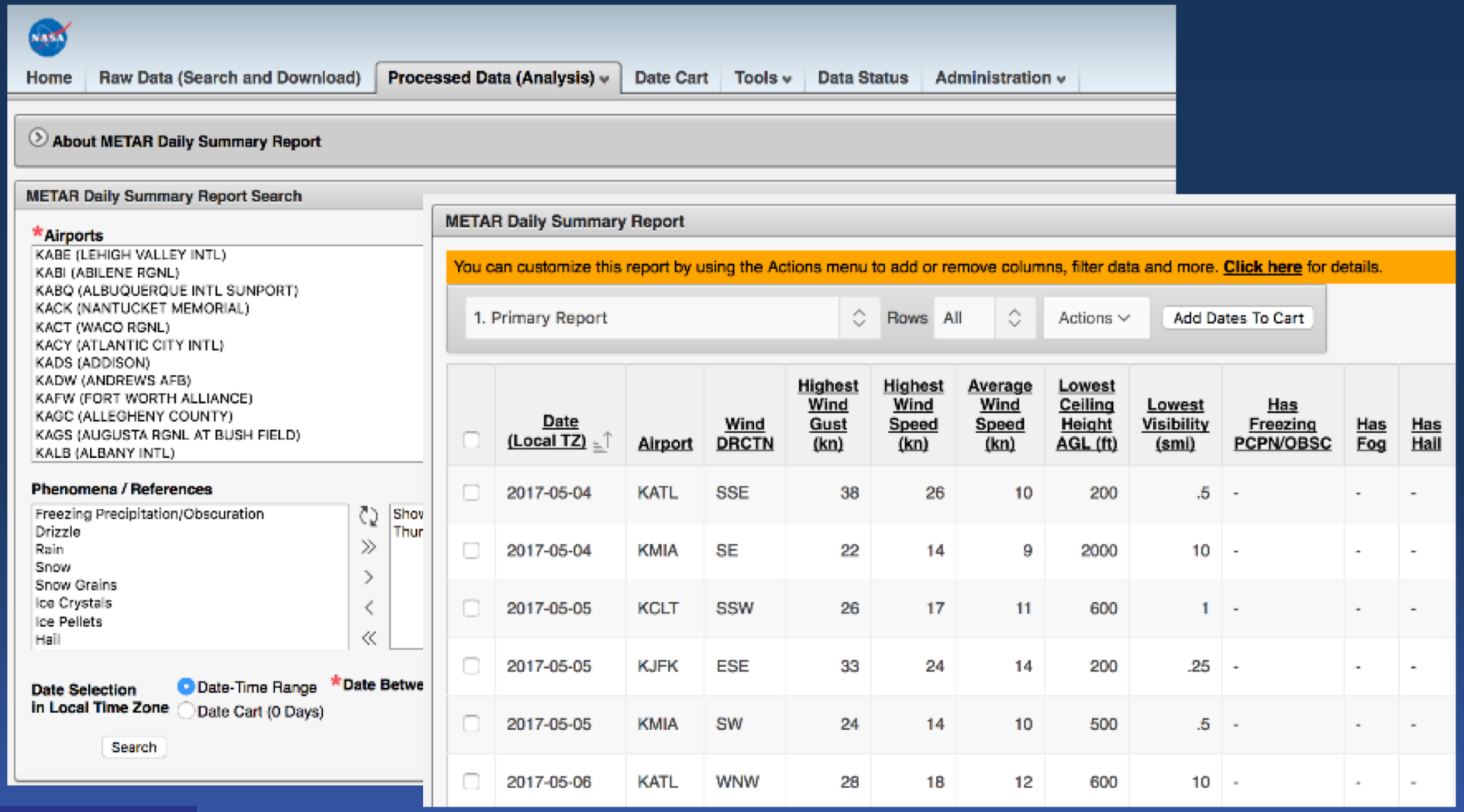
Search and Download

- Search Oracle database across datasets, sub-sets
- Search over dates of interest via range or date cart
- Assess data completeness
- Download selected datasets

Search and Download



Search Parsed Data (METAR)



Date (Local TZ)	Airport	Wind DRCTN	Highest Wind Gust (kn)	Highest Wind Speed (kn)	Average Wind Speed (kn)	Lowest Ceiling Height AGL (ft)	Lowest Visibility (smi)	Has Freezing PCPN/OBSC	Has Fog	Has Hail
2017-05-04	KATL	SSE	38	26	10	200	.5	-	-	-
2017-05-04	KMIA	SE	22	14	9	2000	10	-	-	-
2017-05-05	KCLT	SSW	26	17	11	600	1	-	-	-
2017-05-05	KJFK	ESE	33	24	14	200	.25	-	-	-
2017-05-05	KMIA	SW	24	14	10	500	.5	-	-	-
2017-05-06	KATL	WNW	28	18	12	600	10	-	-	-

Query and Reporting

- Database queries on parsed sources
- Filtering, grouping, charting
- Download results

Sample Chart

OPSNET Delays By Day Report

You can customize this report by using the Actions menu to add or remove columns, filter data and more. Click here for details.

Rows 500 Actions Add Dates To Cart

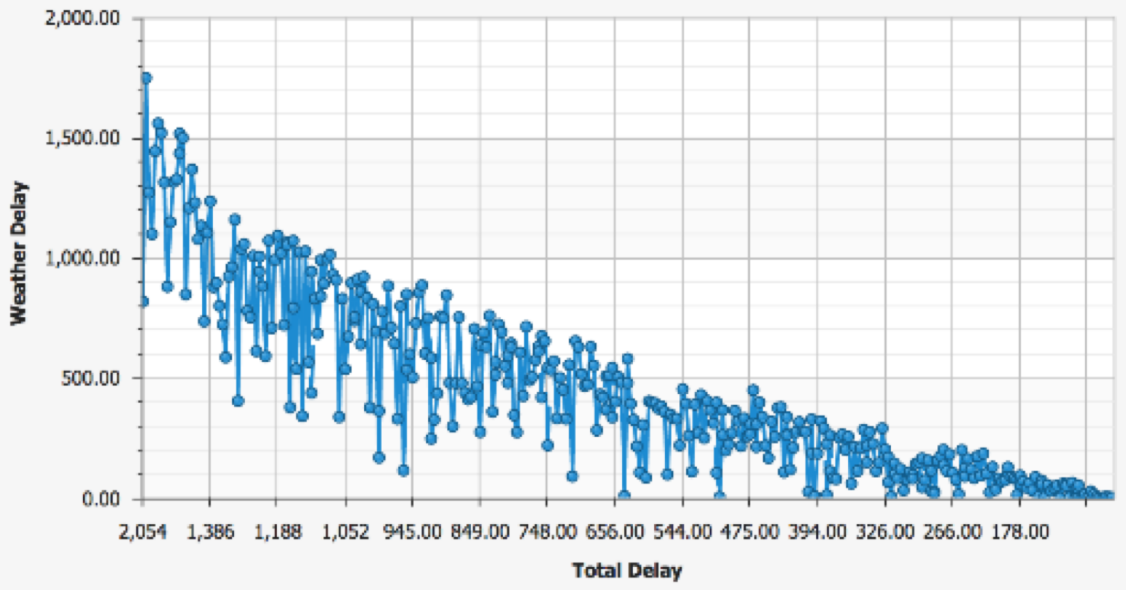
Report Date	Total Ops	Total Delays	TMI To	Dep Delays	Abm Delays
2016-10-28	44,385	2,054	1,656	398	0
2017-01-22	42,908	1,898	1,564	334	0
2017-02-17	53,973	1,886	1,143	743	0
2017-03-02	44,261	1,885	1,293	592	0
2017-01-23	49,041	1,878	1,450	428	0
2017-03-31	43,434	1,806	1,420	386	0
2016-04-07	44,381	1,702	1,481	221	0
2016-09-30	43,783	1,661	1,437	224	0
2017-02-16	57,588	1,646	971	675	0
2016-09-29	44,537	1,637	1,280	357	0
2016-11-30	42,838	1,623	1,346	277	0
2016-11-15	42,831	1,606	1,291	315	0
2016-06-23	45,636	1,597	1,437	160	0

OPSNET Delays By Day Report

You can customize this report by using the Actions menu to add or remove columns, filter data and more. Click here for details.

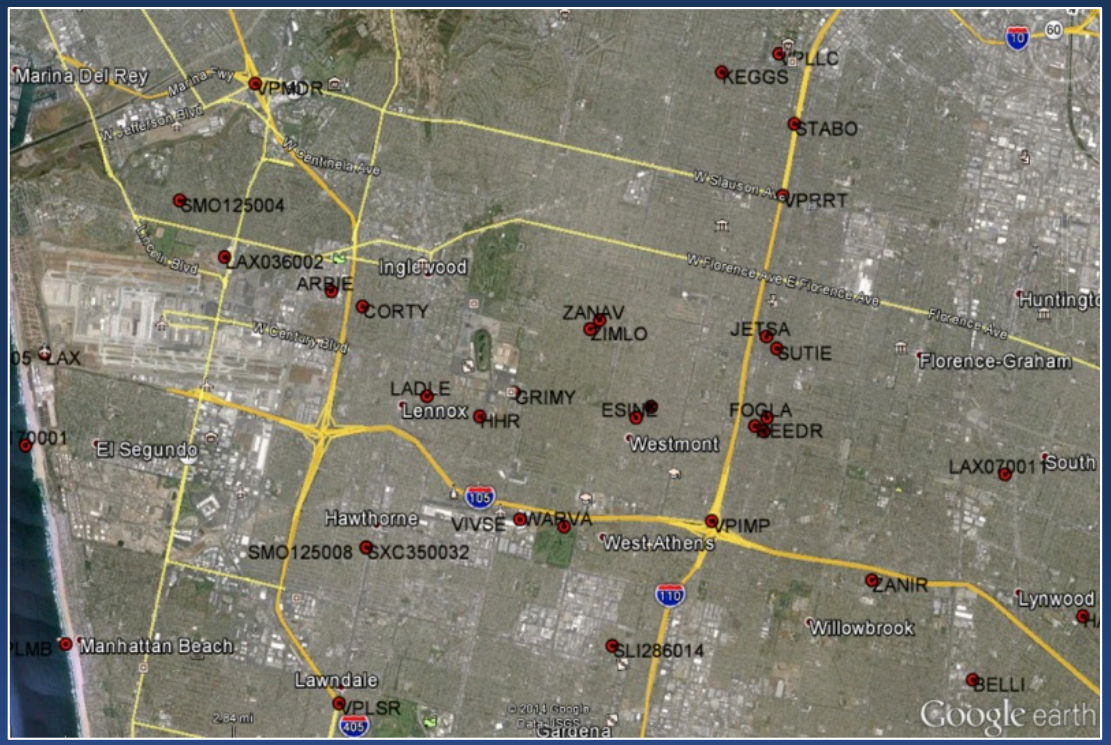
Rows 500 Actions Add Dates To Cart

Edit Chart



Geospatial Service

- Open-source GeoServer
- Airspace features
- Convective weather ‘polygons’
- Query, view, save



Weather Server

- Open-source THREDDS software reads weather datasets (CIWS, RR)
- WMS query, visualization, export