SAND2019-10607C

Firebird

Integrated Multi-Phenomenology Data and Analytics Platform

Elaine Martinez, Uen-Tao Wang

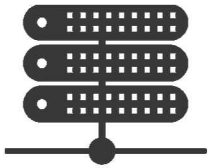
Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

Topics

- Big data
  - Characteristics
  - Problems
  - Use Cases
- Current big data ecosystem
- Firebird
  - Architecture
  - Where is it being used
- Lessons Learned
- Design efforts for Fb 2.0

Big Data Characteristics



Volume

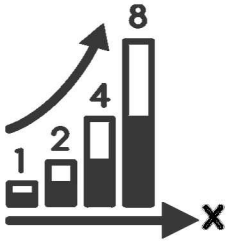
How big?

Large volumes of data associated with raw sensor data, information from data processing systems

Variety

How Different?

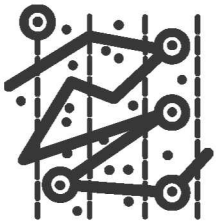
Temporal, spatial, spectral data in many formats and standards



Velocity

How fast?

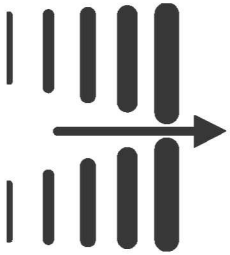
Real-time with a constant flow of data or sensors and sensor processing systems



Variability

How data is changing?

Customer requirements, changes in standards, new systems



Veracity

How trustworthy?

Different accuracies and trustworthiness we must be able to handle

Characteristics require a scalable architecture for efficient storage, manipulation, and analysis

http://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.1500-1.pdf

Data Management

Common problems faced by data professionals across industry sectors

- Inability to handle the speed and volume of **multi-source** data
- Inability to find a **single technological** solution to collect, store, and organize data from disparate sources
- Inability to handle big data projects with a **single database** technology
- The increased adoption of **cloud platforms** and cloud infrastructures has raised concerns regarding data security

https://www.dataversity.net/data-architecture-trends-in-2019/

Why a big data architecture?

To enable applications and customers that require the integration of multiple heterogeneous data collections

	Use Case		Goal
Intelligence	Ingest/accept data from a wide range of sensors and sources across intelligence disciplines	→	Provide automated alerts to Analysts, Warfighters, Commanders, and Leadership based on incoming intelligence data
Geospatial	Data sources are increasing and complex analysis and visualization is growing. Today's systems often contain trillions of geospatial objects and need to visualize and interact with millions of objects	→	Support large scale geospatial data analysis and visualization
Climate	Researchers, weather forecasters, instrument teams need to access data across multiple datasets to compare measurements, models, calibrate instruments and correlate across parameters	→	Automate the discovery of diverse data, decrease data transfer latency, and meet customizable criteria based on data content, data quality, metadata, and production.

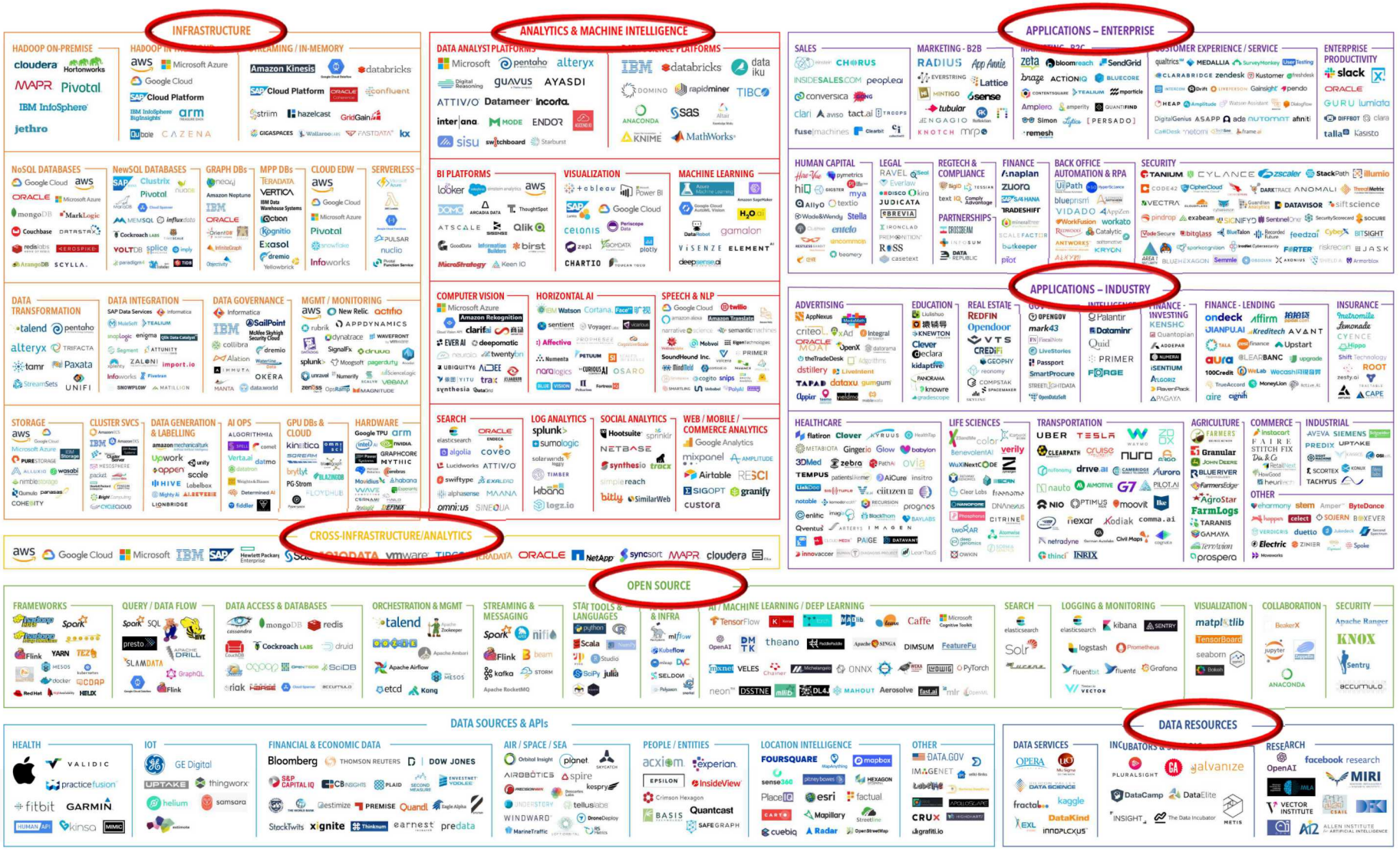
Big Data Ecosystem

What's Needed

What's Enabled

The Ecosystem is Maturing

Data & AI Landscape 2019



FIRSTMARK

Big Data success is not about implementing one piece of technology (like Hadoop or anything else), but instead requires putting together an assembly line of technologies, people and processes.

Requires Specialization in Enabling Data and Information Exposure

- Data engineers must understand **cross-functional use cases** and work with domain specialists to perform **requirements elicitation** and **data understanding**
- Understand how to:
  - Optimize for data analysis on **important** data
  - Allow reach back to **ALL** data
- Expertise in diverse computer science technologies and languages
  - Maintain awareness of **emerging paradigms** and approaches in order to be able to identify and apply the **right tool for the right job**
- Able to design the right architecture for the right need
  - Can articulate pro's and con's of various platforms and solutions

Skilled architects with cross-functional (full stack competencies) and cross-domain know-how

https://martinfowler.com/articles/data-monolith-to-mesh.html#DomainDataCross-functionalTeams

Enables Data Analytics

- Facilitate the exposure of **multiple** data sources
- Facilitate user driven integration of **multi-phenomenology** data sources
  - Multi-Phenomenology data sources are made available
  - User selects data sources needed to perform analysis
- Architecture features **composable/modular** design
  - Technical components are interchangeable based on requirements of customer

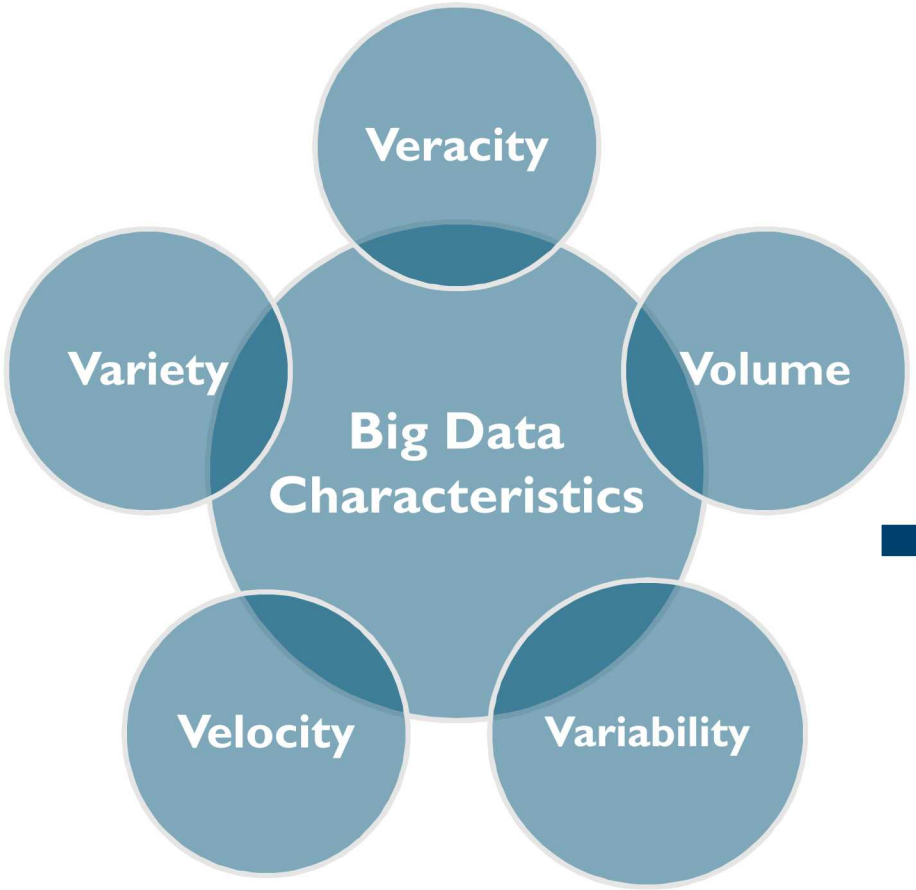
What have we built?

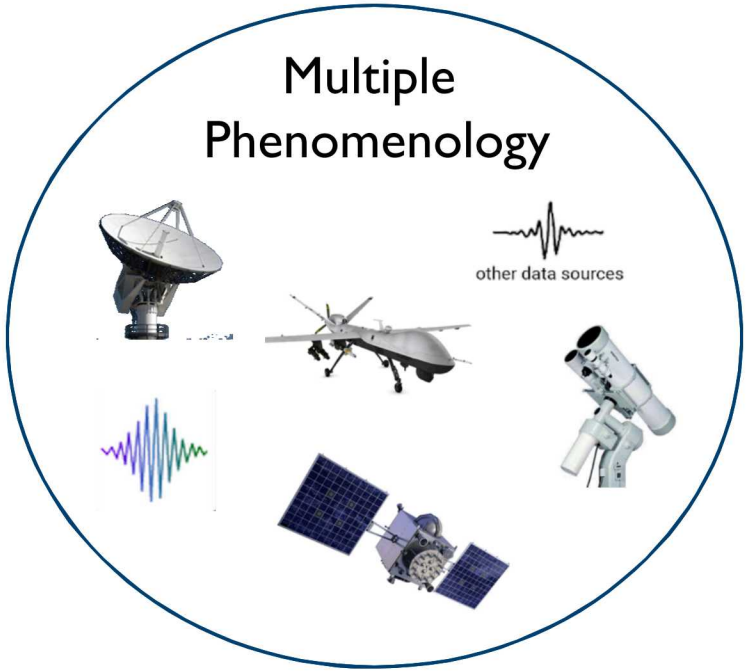
FIREBIRD

- Functional Stack
- Architecture

How to address big data use cases?

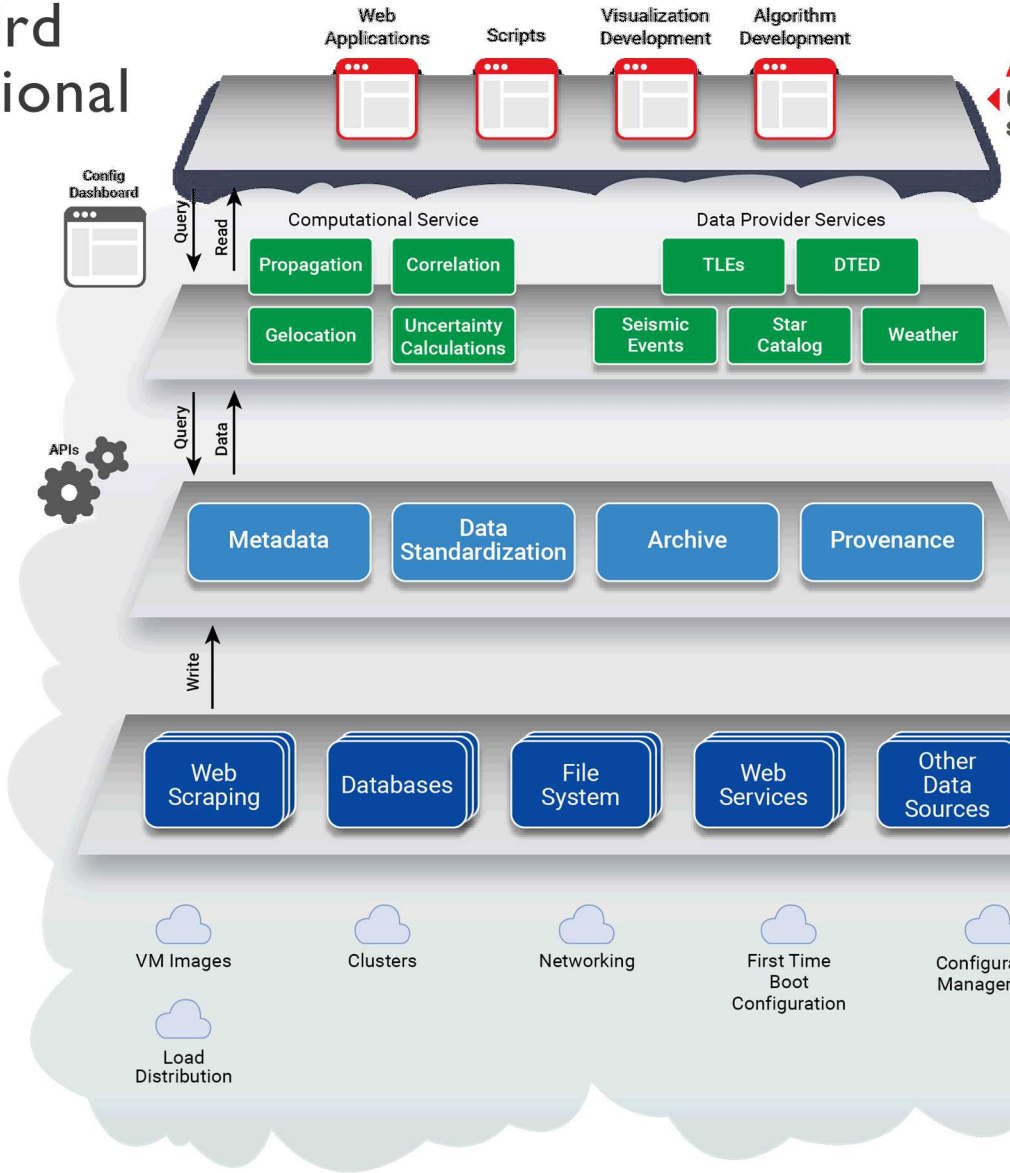
How is analysis enabled across **multiple characteristics** and **multiple phenomenology** data sets?





Design an architecture that is scalable enough to incorporate big data characteristics and is flexible enough to integrate multi-phenomenology sets of disparate data

Firebird Functional Stack



APPLICATIONS AND VISUALIZATIONS
Orchestrate reusable elements in various ways to solve a wide range of problems.

SERVICE LAYER
Technical components which combine together to give system flexibility & reliability. Orchestration of data and computational services provide dynamic addition/removal of functionality as new capabilities are developed.

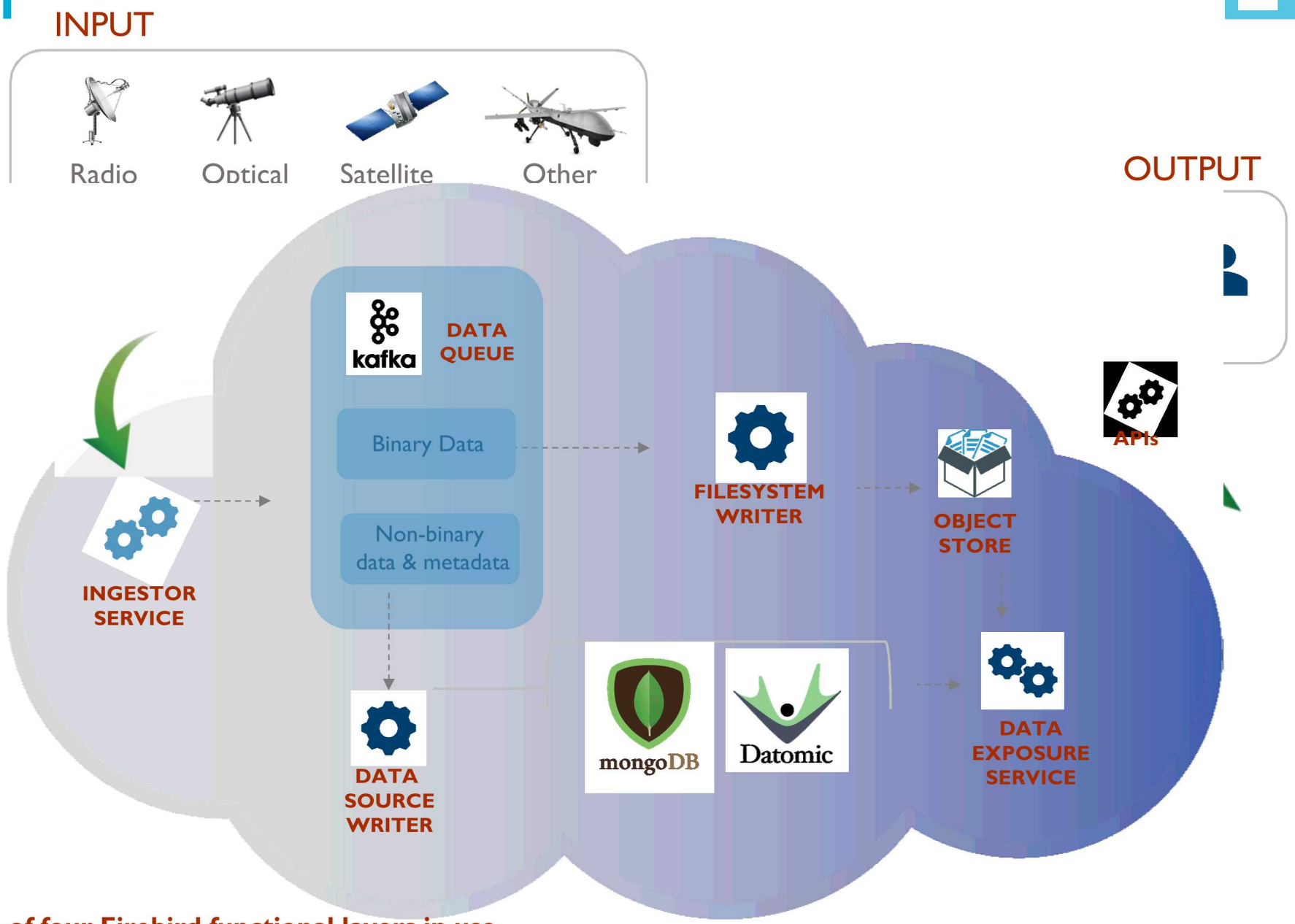
DATA LAYER
Ingest, standardize and archive data from a variety of sources for consumption by users.

EXTERNAL DATA SOURCES
Sources can be raw, unstructured or structured data.

CLOUD INFRASTRUCTURE
Scalable and reliable cloud platform which includes rapid deployment techniques.

Firebird implements a reusable **multi-layer** architecture designed to enable user communities to use a **diverse sets of data** that exists in **various formats** from a **variety of sources**; and better facilitate analysis, collaboration and sharing.

Firebird Architecture (project use case)



Two of four Firebird functional layers in use

Shapes and Luggage

Shape

A data shape enables discovery, is queryable and allows the tracking of luggage

- Shapes are how **standardization** across disparate data sets is managed
- Allows for **complex operations** on **diverse data sets**
- This approach **combines the power** of
  - Set schema in something like a data warehouse
  - Flexibility and retention of all data (even data that doesn't conform to a schema) of a data lake

Luggage

Luggage (untransformed data), is data that is being pulled along with the data in a shape

- **Untransformed** original data
- The importance of luggage is that the **data is unchanged**
- Search across luggage fields is ad hoc, and potentially slow
- This data should be self-describing so that when users get it out of Firebird, users know how to process it
- All **original data is kept**, including data that is also represented in shape fields

Example Seismic Shapes

Seismic Station

- name
- latitude
- longitude
- elevation
- on-time
- off-time
- [networks]
- [sites]
- luggage

Space Ground Station

- name
- latitude
- longitude
- elevation
- on-time
- off-time
- [constellation]
- [frequency]
- luggage

Data engineers and users work together to map data into standardized shapes

Lessons Learned

Lessons Learned

Data

- Data maturity is important
  - Can't develop a Firebird for customer when data format and use cases were being determined
- Don't ingest data from original source to Firebird
  - Goal is **not** to become a data warehouse
  - If needed, should establish "contract" with customer to reach back into their data source
- Very hard to establish a one-size fits all process

Security & Classification

- Some customers didn't allow us to show/use data
- How to combine data from different classification systems
- Classification guides are lacking or absent entirely

Customer

- Needed time from SME to understand use cases and shape design

Lessons Learned

Technology

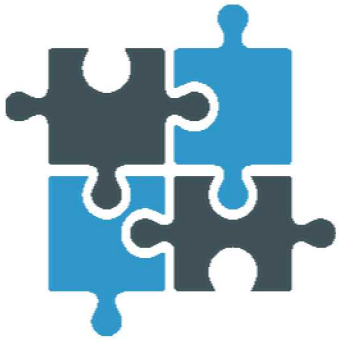
- Persistent Storage Solutions: Initial prototype used NFS; slow
- Docker Compose is fine for single system deployment, but non-trivial work required to use Kubernetes or Swarm
- Strictly using GraphQL made it hard to do blob retrieval as it doesn't support sending binaries out of the box
- Having separate solutions for storing binary data vs non-binary data made it difficult to sync
- System was designed using several services, but in early development stages any change was a breaking change

Overarching

- We were trying to anticipate the needs – make as powerful as possible
- The more generic – broader scope we make it, the less powerful the system may be
- The more complex, the more powerful (query ability), limited scope

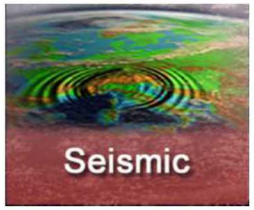
Successful in:

- Plug and play design & technology – right tool for the right job
- Design for scalability
- Extensibility – design for current and future use cases
- Re-usability – design to be data and mission agnostic

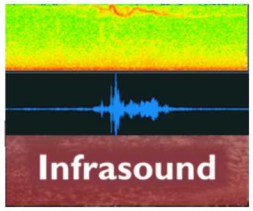


Current R&D Customers and Data Domains

Existing multi-phenomenology data brought into Firebird



Coming soon….



Consulting for:

Thoughts on Firebird 2.0

GOALS

- **Embrace Data Virtualization - not to read in copies of existing data sets into a location we manage**
- **Identify what's already been done to help identify what WE need to do**

- Investigate existing technologies for enabling the data mesh/data virtualization
  - Facilitates access to disparate data sources and neither owns data nor imposes constraints
- Investigation of current landscape
  - Understand what's out there
  - Understand how these products can be leveraged, possibly in concert with one another
  - Understand how these products can be extended
  - Try to use newer tools with existing data sets
- Investigate solutions for reconciling disparate querying capability of data sources
  - e.g. Can perform geospatial queries in PostGIS but not in a file system

Won't expect to find a one-size-fits all solution

Questions