Exceptional service in the national interest

Sandia National Laboratories

A Platform for Provisioning Integrated Data and Visualization Capabilities

Presented to SATURN in May 2016

Gerry Giese, Sandia National Laboratories

U.S. DEPARTMENT OF ENERGY

National Nuclear Security Administration

Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000. SAND NO. SAND2016-3884 C

The Business Problem

Sandia National Laboratories

The creation of the Analytics for Sandia Knowledge (ASK) platform was precipitated by a data study that was severely bogged down in the data collection/preparation phase. We found that we were not alone in this difficulty, both within our organization and in many other companies.

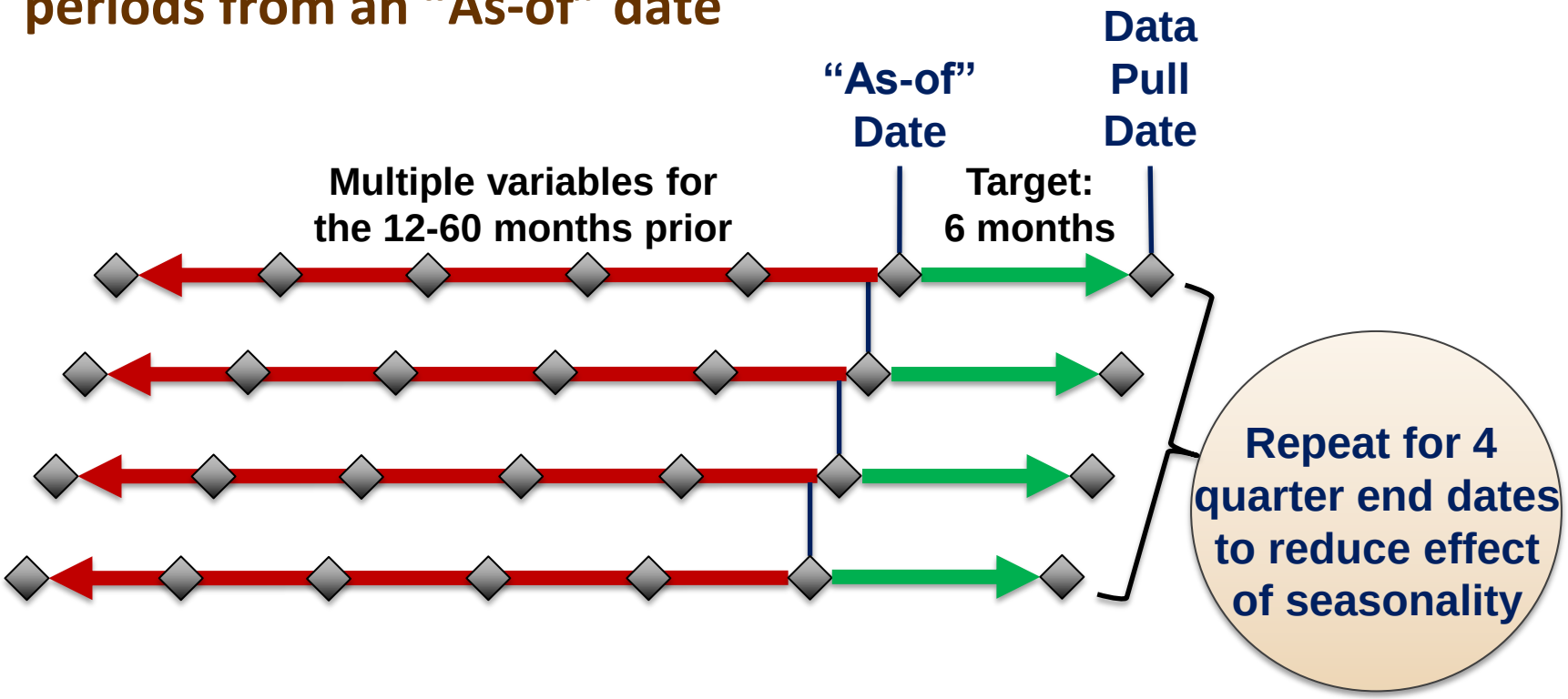
"Data scientists, according to interviews and expert estimates, spend from **50 percent to 80 percent of their time** mired in this more mundane labor of collecting and preparing unruly digital data, before it can be explored for useful nuggets."

Lohr, S. (2014, August 18). For Big-Data Scientists, 'Janitor Work' Is Key Hurdle to Insights. *New York Times*. Retrieved from http://www.nytimes.com/2014/08/18/technology/for-big-data-scientists-hurdle-to-insights-is-janitor-work.html

Example Data Study

Sandia National Laboratories

Data points are collected and summarized over multiple time periods from an "As-of" date



This is just for one data point! Need to integrate many data points that may not align well, may not have history, …

Example Data Study

Sandia National Laboratories

Data was gathered (slowly) from multiple data sources

- Human Resources systems
- IT systems management records
- Safety & Health records
- Training records
- And many more…

No Corporate Process

- Commonly heard: *That's MY data! What are you going to do with it?*
- Had to brief 4 Directors, 1 VP, and 3 Lawyers
- As a result, ASK is implementing a corporate data governance process to enable faster/safer data collection for future studies

The Business Problem, Revisited

Sandia National Laboratories

The challenges faced when performing data studies – data collection, integration, quality, etc. – are complex and time-consuming, limiting the number of studies performed (and therefore questions asked).

When data analysts provide answers, the data they used is no longer current and the results represent a look back at a point in time, resulting in significant additional effort to "refresh" the results.

Primary Needs:

- **Finding Data** – analysts were starved for data
- **Integrating Data** – data sets need better 'alignment'
- **Historical Data** – need to recreate "point in time" snapshots
- **Provisioning Data** – needs to be much, *much* faster
- **Visualizing Data** – more flexibility and capabilities, rapid development
- **Data Confidence** – improve quality and pedigree/lineage
- **Data Governance** – ensure safety, privacy, security

System Plan

Sandia National Laboratories

ASK Goals

- Accelerate time-to-results for building analytic models
- Explore integrated data sets to derive and discover knowledge
- Expose data and analytic results and interactive visualizations of them to users

Primary Constraints

- The platform will only integrate and transform existing data found in source systems, or provide warehousing of existing data in source systems in order to support historical views
- Don’t propagate Personally Identifiable Information (PII) data into new repositories
- When you do expose PII (virtually) to those with authorization, “anonymize” it as close to the original source as possible
- Web-based interface, support mobile users

Architecture Plan

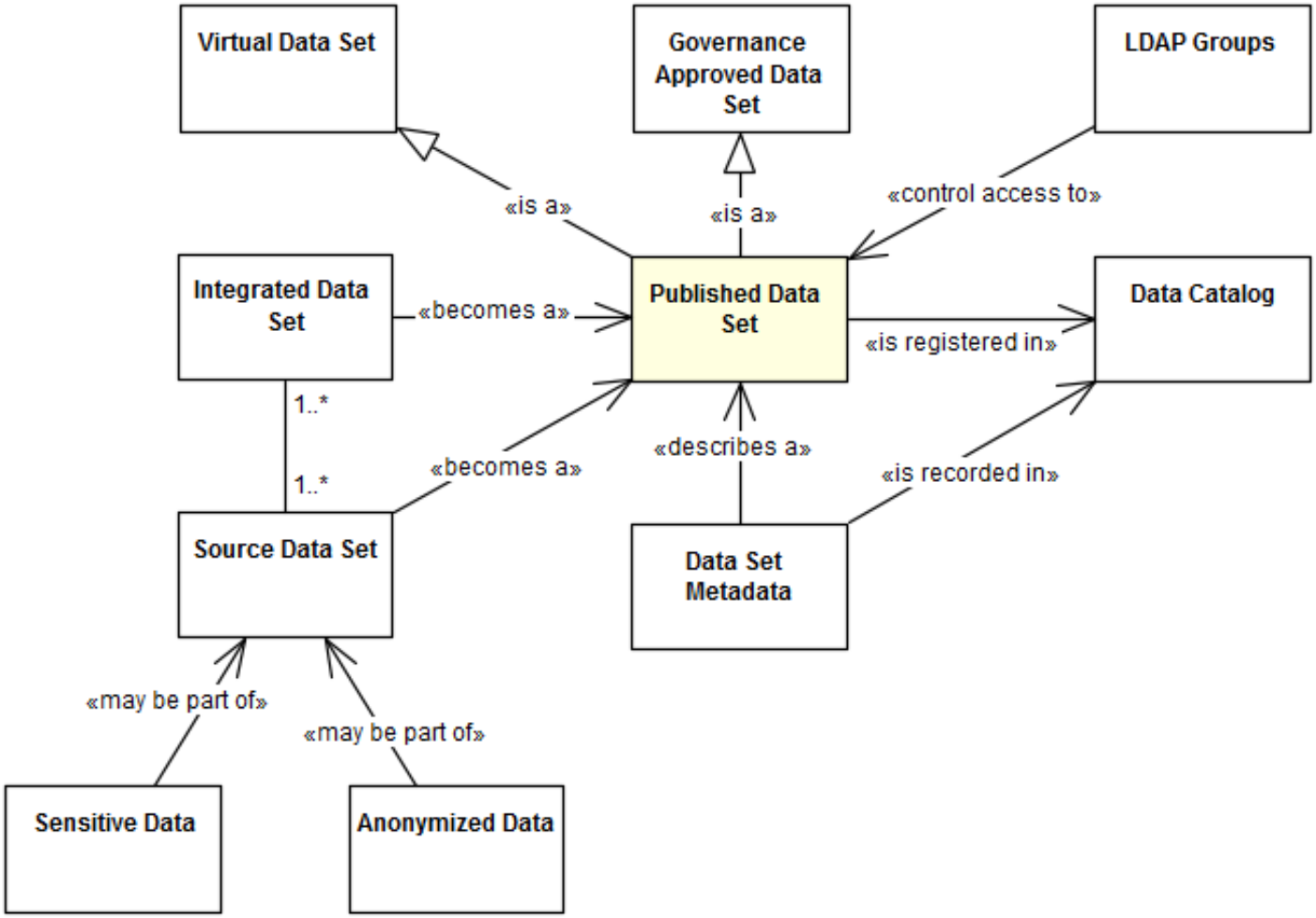
Sandia National Laboratories

System stakeholders and the system team developed a conceptual data model, roughed out a process, developed and prioritized quality attributes, identified system constraints, captured relevant target architecture features in our program area, then derived a high-level system target architecture.

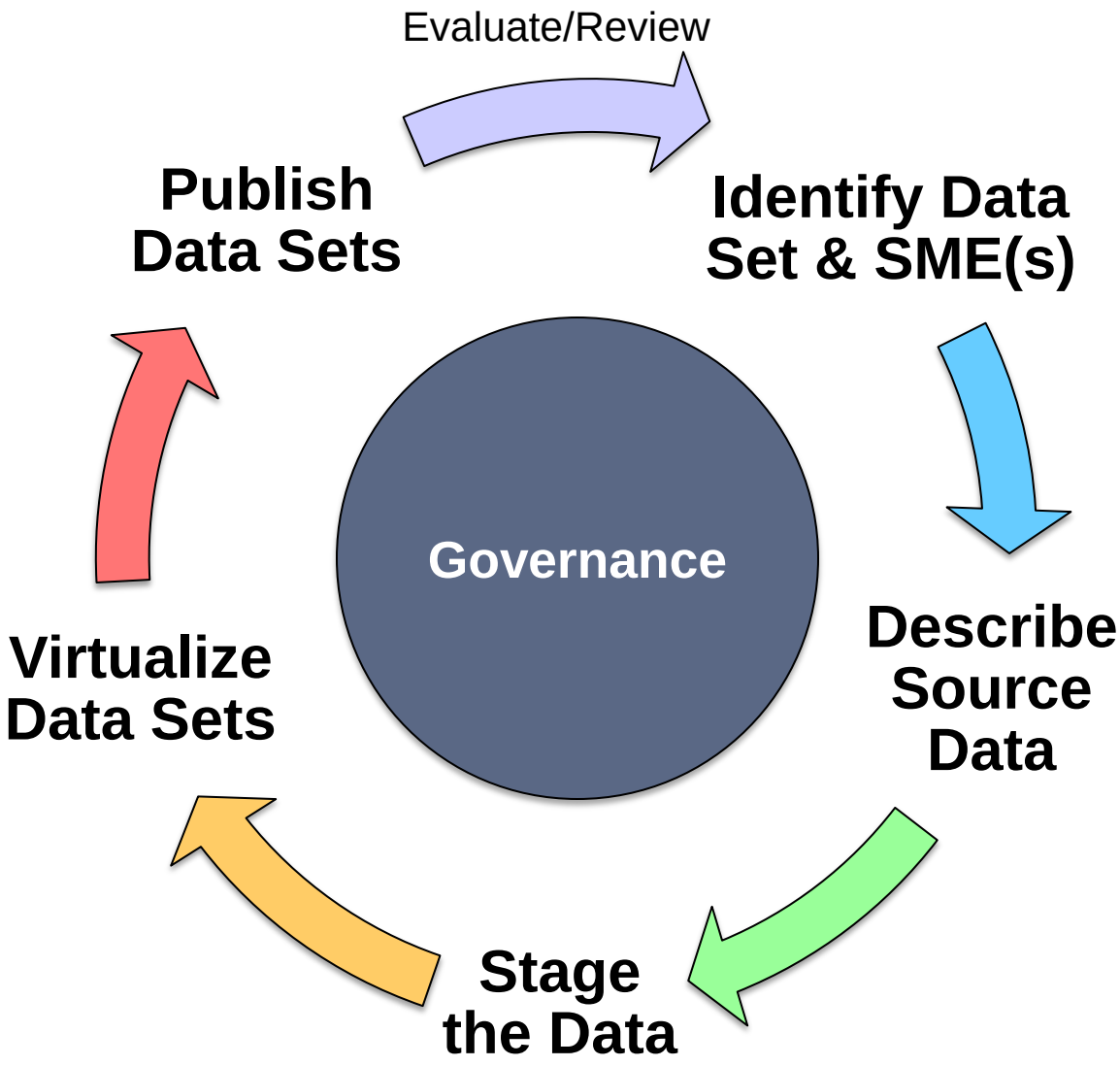
Quality Attributes:

- **Privacy** – legal and ethical assurance
- **Security** – prevent unauthorized access or data breaches
- **Usability** – support existing tools; avoid complexity
- **Accuracy** – use the right data; repeatable results
- **Performance** – users shouldn't have to wait
- **Flexibility** – support multiple access methods
- **Availability** – maximize system uptime, on a budget…

Conceptual Data Model



Simplified Process

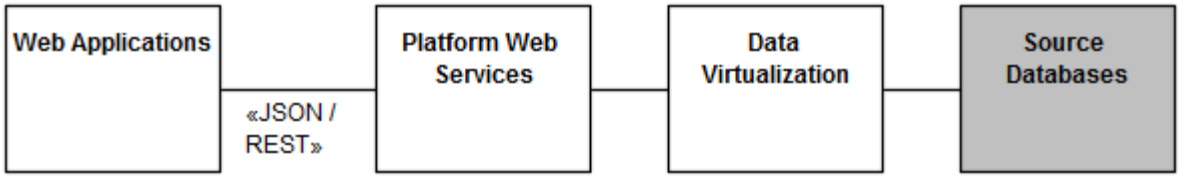


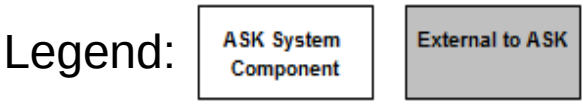
Initial Architecture

Sandia National Laboratories

Given what we knew to start with, we began with a modern web application pattern of a JavaScript client application, RESTful JSON web services, and a data tier.

Unique to our situation, we understood that we would have read-only access to source data in disparate forms and technologies, necessitating an additional data layer to provide transformations and connectivity. Data Virtualization had been on our radar for a while and appeared to be a good fit.

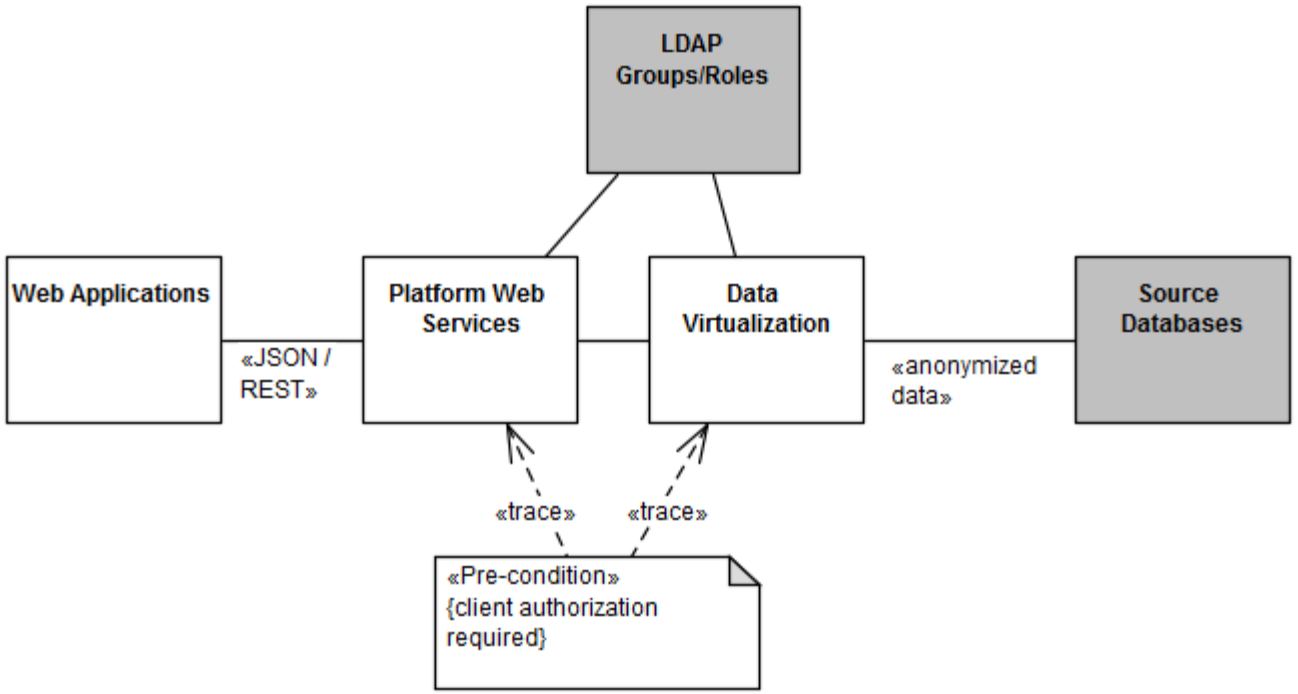




Architecture QA: Privacy

Sandia National Laboratories

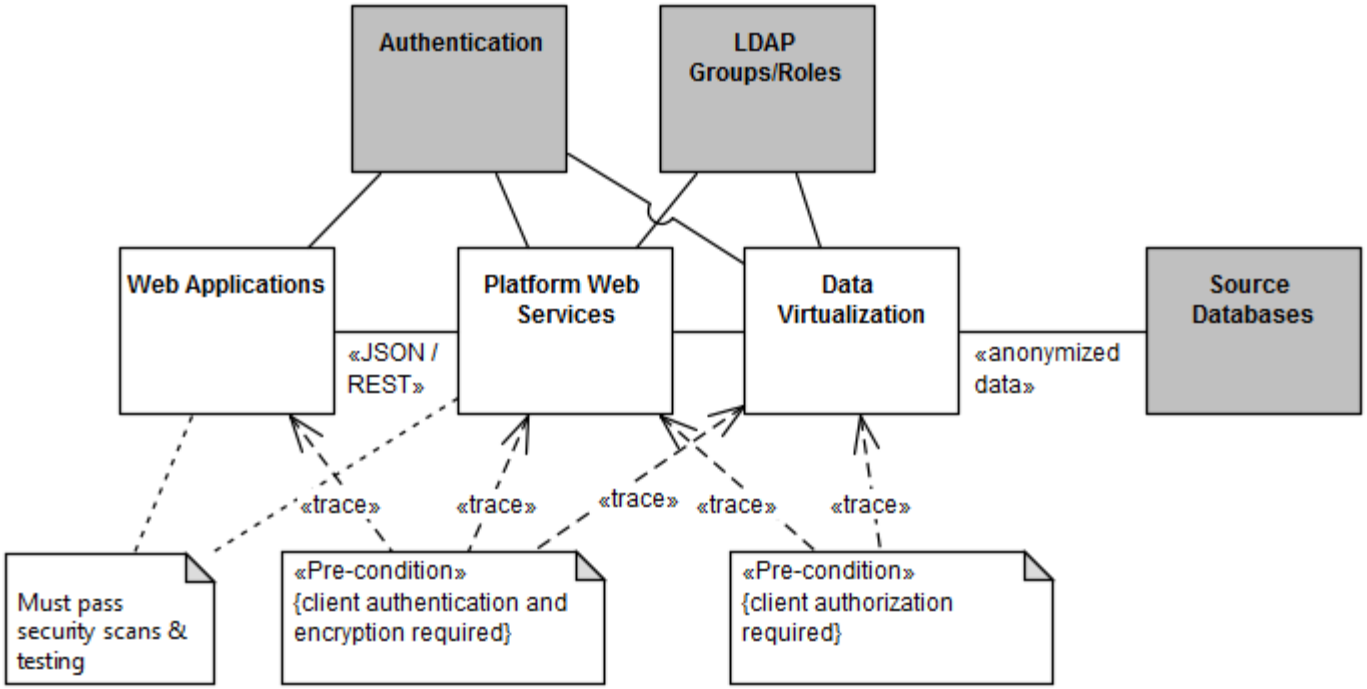
HR data entities were an early target of data provisioning in the platform, and the chief concern we received around publishing this data is that privacy is protected and only users with a need to know are granted access to it.



Architecture QA: Security

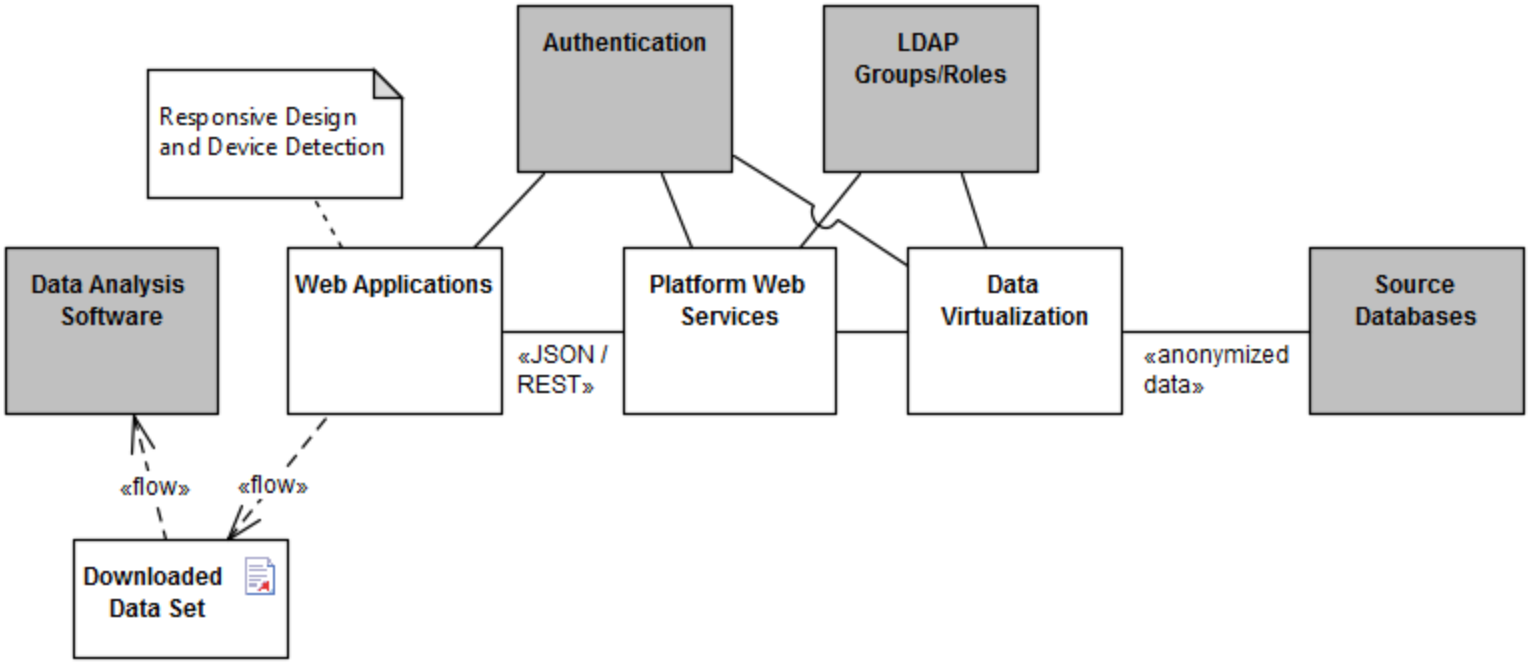
Sandia National Laboratories

Given the type of work Sandia performs as a national laboratory, security is a high priority for any system we build. The distributed processing and expectation of multiple platform client types meant each layer and component must be secured independently.

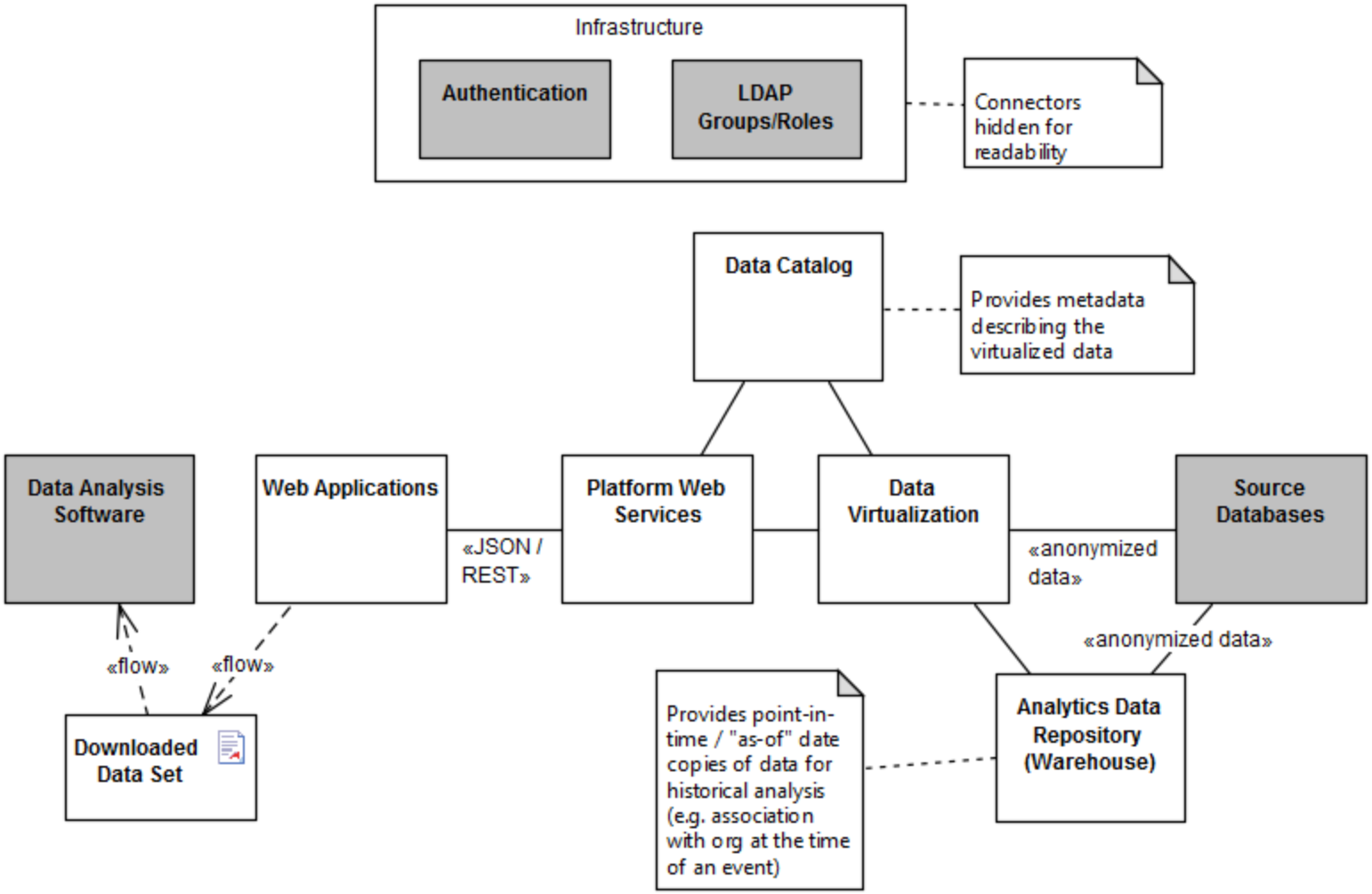


Architecture QA: Usability

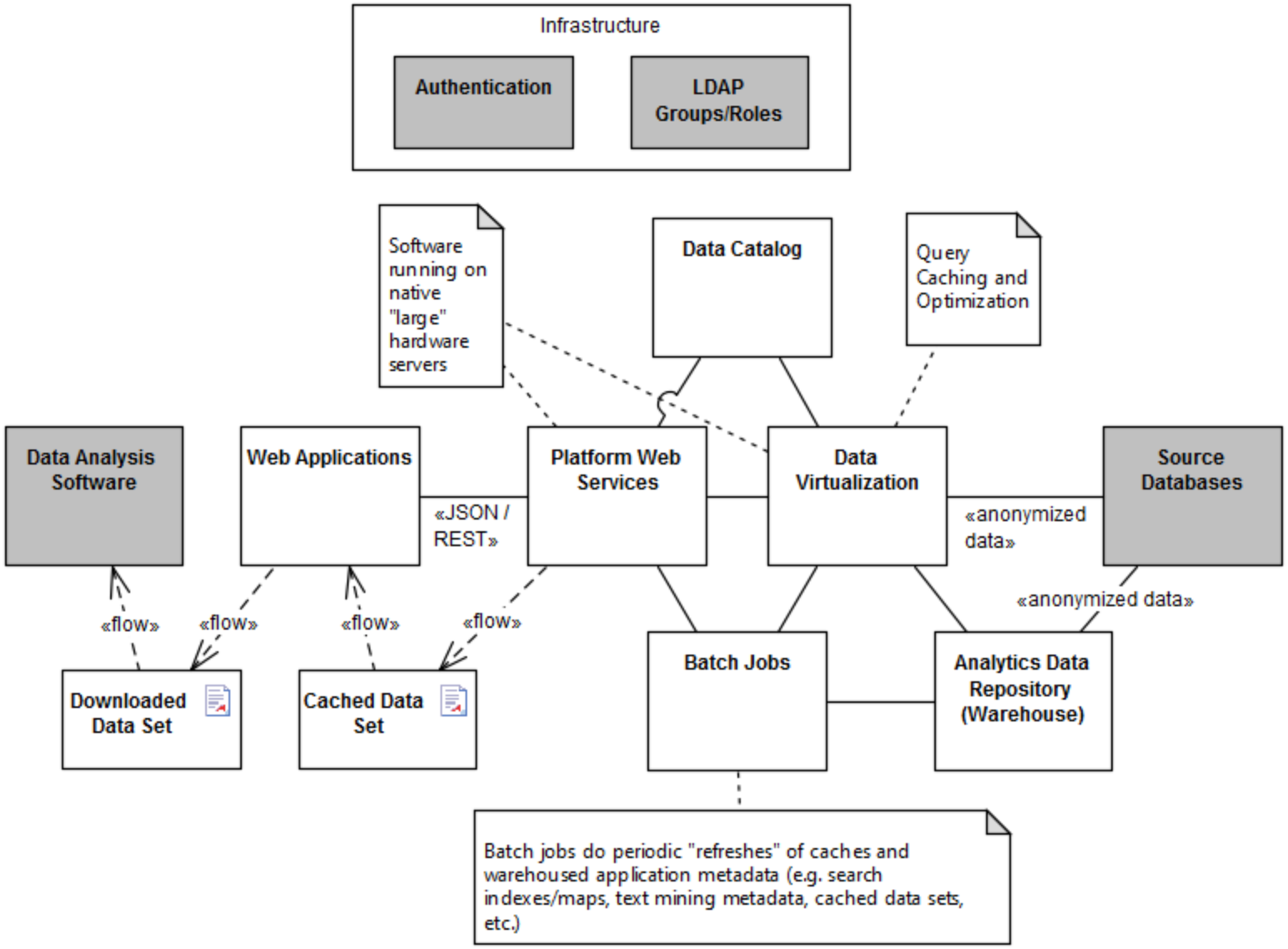
Most of our users will use a web interface to interact with the platform and applications. With users ranging from executives to staff, we needed to provide an interface capable of running well on multiple device types. Additionally, data analysts each have their own favored toolset, so providing a data download capability was important.



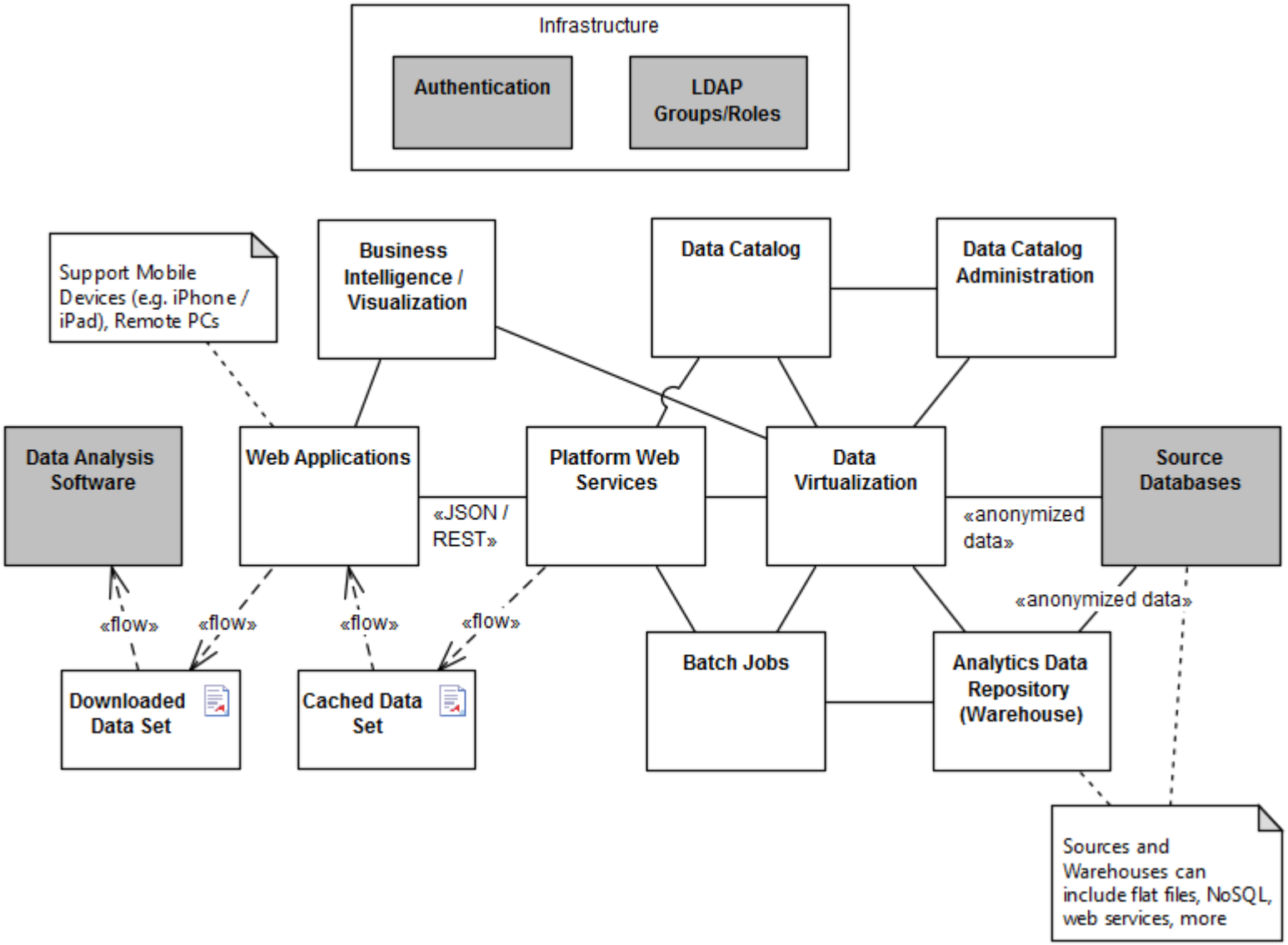
Architecture QA: Accuracy



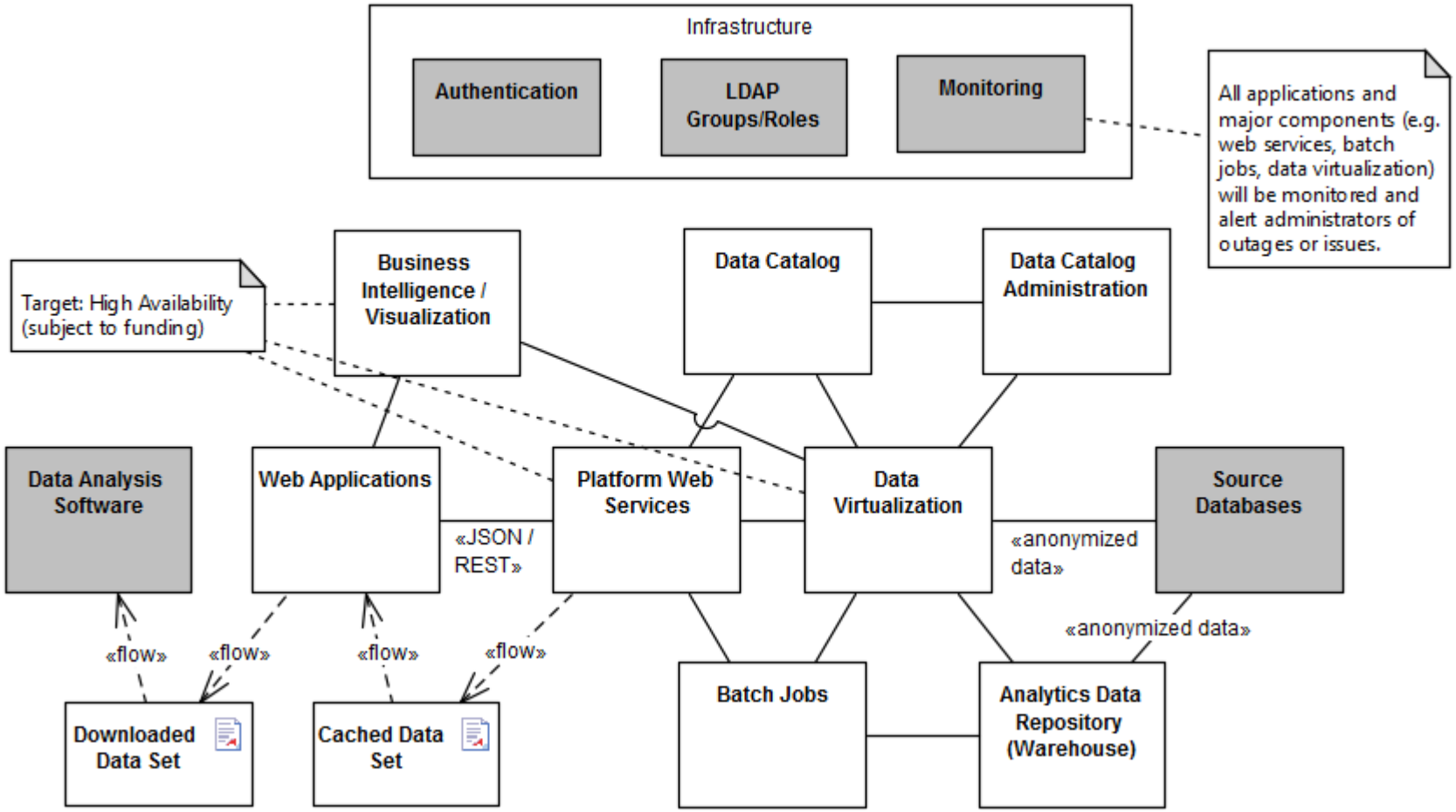
Architecture QA: Performance



Architecture QA: Flexibility



Architecture QA: Availability



Key Features/Takeaways

Sandia National Laboratories

Data Governance

- Engage SMEs to identify and describe data sets; guide them toward compatible definitions that integrate well
- Build a data catalog to capture business-friendly data definitions and to track data lineage (note: your data toolset may provide reports to do this for you, up to a point)
- Map out your roles and access rules out and simulate with users and stakeholders to tease out issues, but always aim for simplest solution that works – may not be the status quo
- Give users access! But as data sensitivity grows, ensure they understand how to protect it. Also “trust but verify” with audit logging, and demonstrate an audit to stakeholders.

Key Features/Takeaways

Sandia National Laboratories

Data Foundation

- Identify data sets that can't recreate a point in time, take regular snapshots
- Use data virtualization to complement ETL. Virtual layers in your data pipeline, from source to final product, will reduce re-work and enforce separation of concerns
- Data virtualization (abstraction) lets you expose business-friendly data naming to end users
- Provide stand-alone data entity products (e.g. person, location, event, etc.) as a baseline, then build virtual integrated data sets from them as use cases/patterns arise

Key Features/Takeaways

Sandia National Laboratories

Data Visualization

- Use a client-side web application platform to integrate multiple visualization technologies into a (mostly) consistent user experience where users can quickly find the visual they need
- Encourage analysts to vet and share results, possibly in different forms for mgt. vs. staff
- Interactive visualizations are much more engaging to users
- A web-based self-service visualization tool is in development, using the business-friendly data provided by our platform
- User experience, user experience, user experience… just because it's visual does not mean it's intuitive!
  - Shaping the data before consumption will improve results