BOOZ | ALLEN | HAMILTON

Spark + AI Helps the FDA Protect the Nation

Kun Ei Kang
Booz | Allen | Hamilton
Chief Innovation Architect

Jonathan Chu
Booz | Allen | Hamilton
Chief Technologist

June 5, 2018

www.boozallen.com

AGENDA


OPEN DATA PLATFORM
Booz Allen's ODP
Our FDA Solution Journey
About the FDA
Key Findings
1
2
3
4

FDA AND YOU

- The FDA oversees products which account for **20 cents of every dollar spent** by consumers
- **Foreign production** of FDA-regulated goods and materials has **exploded** over the last decade
- FDA-regulated food and medical products originate from more than:
  - **150 countries**
  - **130,000 importers**
  - **300,000 foreign facilities**

FDA IMPORTS PROGRAM

Responsibilities:

- **Protect public health** by electronically screening all FDA regulated product imports
- **Determine and stop** product line/shipments if they **pose significant risk**
- **Obtain samples** of products for further laboratory screening as needed

High Volume

~40 Billion shipments in 2017 with continuous annual increase of 5-10%

Performance Insight

Sheer volume of data led to **increased labor to assess** performance of electronic screening rules

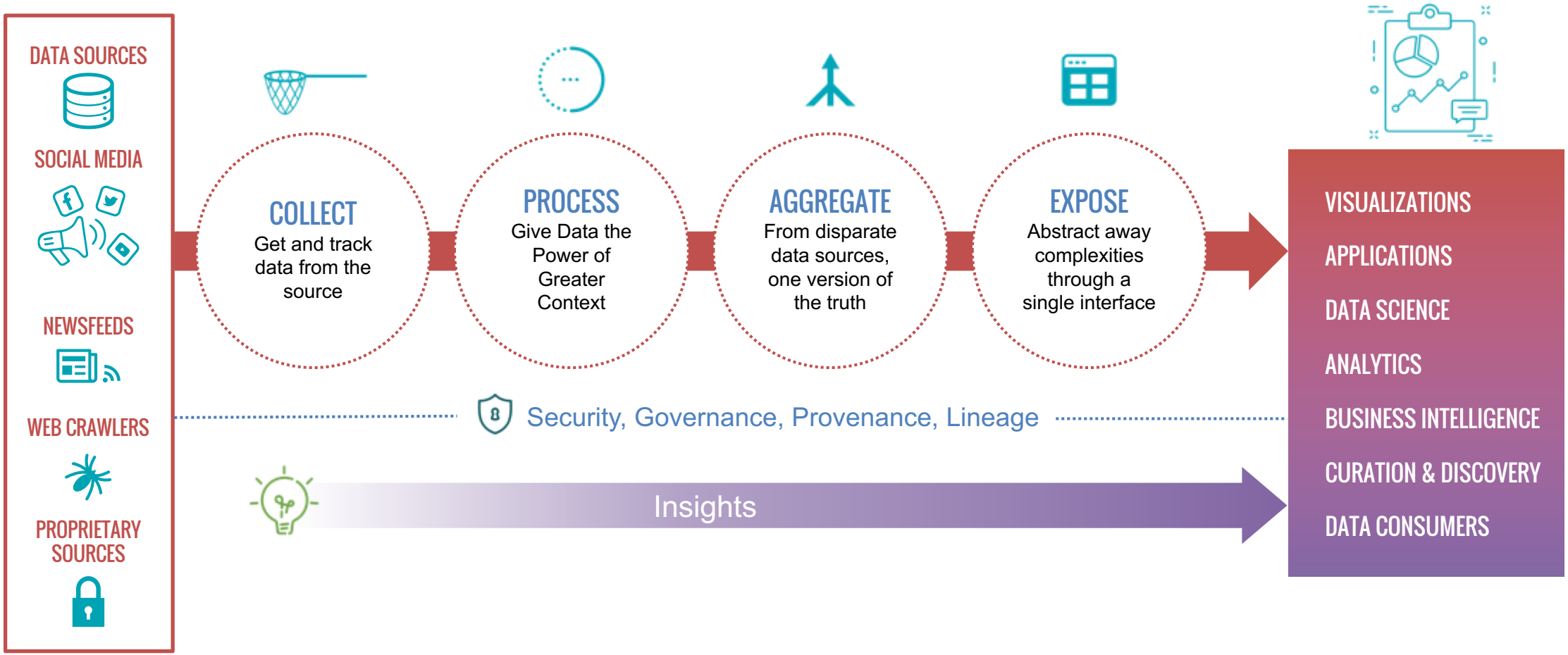
Change Management

Need for more robust capabilities to evaluate proposed changes to screening criteria for imports

Booz | Allen | Hamilton

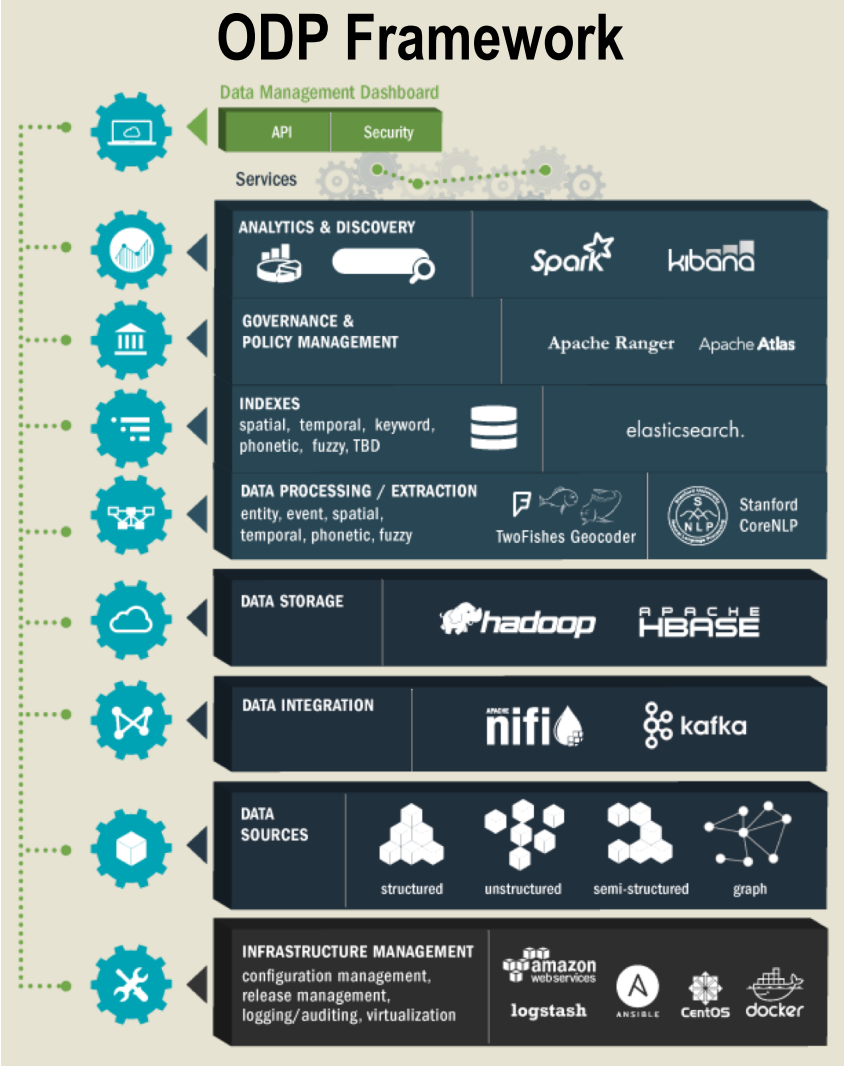
CONSULTING | ANALYTICS | DIGITAL SOLUTIONS | ENGINEERING | CYBER

ODP TURNS DISPARATE DATA INTO VALUE ADDED SOLUTION FOR OUR CLIENTS



ODP CORE TECHNOLOGY STACK

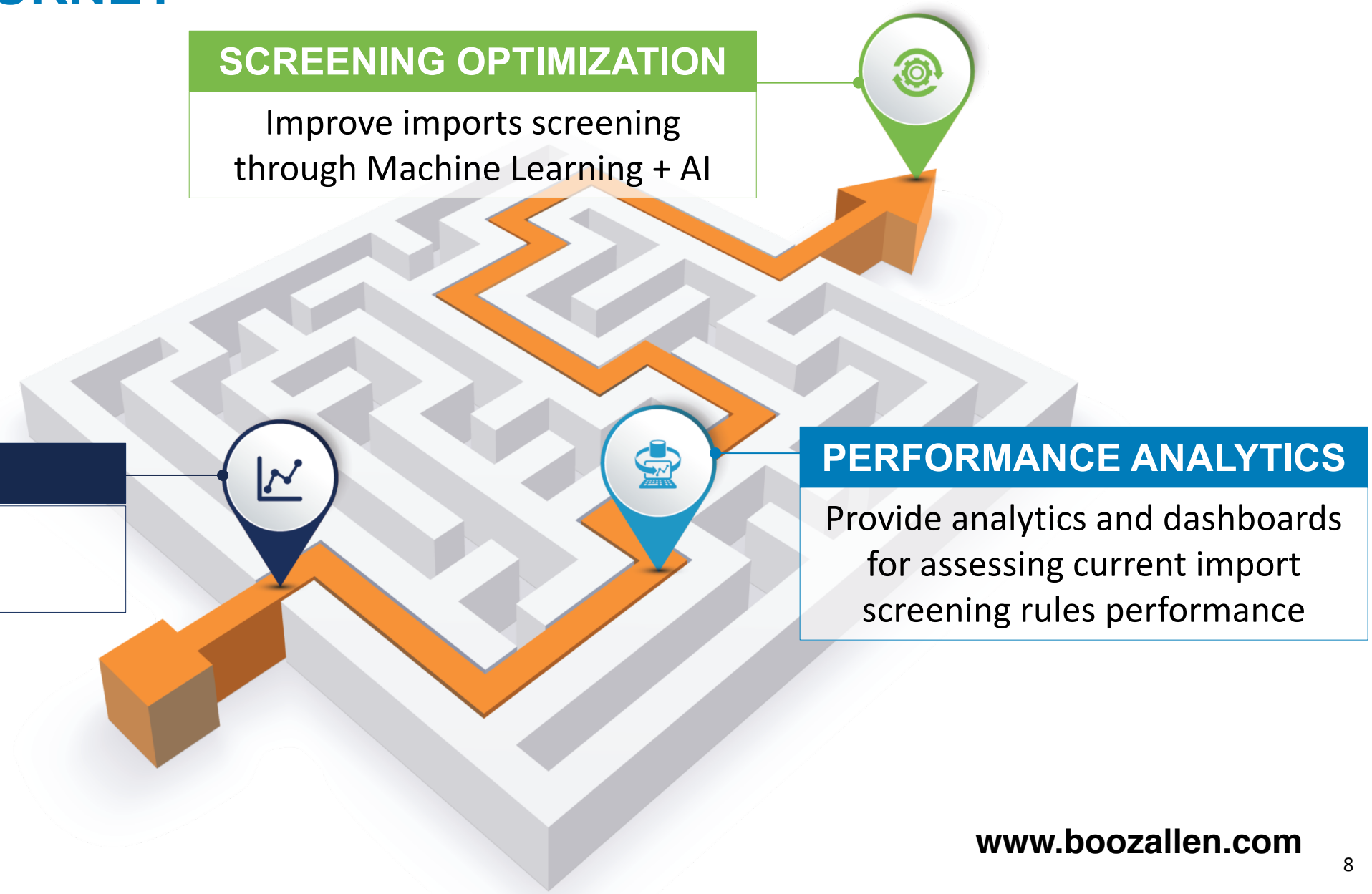
- Best-of-breed open source technologies chosen, configured and integrated
- Containerized data ingest and processing pipeline Apache Spark and Docker
- Automated deployment into Amazon Web Services (AWS) and Azure
- Ability to swap in-and-out technologies based on use case, as well as tailor deployments based on use case (e.g. Elasticsearch vs Solr; deploy search capabilities without Hadoop ecosystem)
- Data Management dashboard for job management, data tracking lineage/provenance, metadata management, and governance
- Analytics platform that democratizes data science and enables analytic decision making for anyone (of any skill set) in an organization



SOLUTION JOURNEY

SCREENING OPTIMIZATION

Improve imports screening through Machine Learning + AI



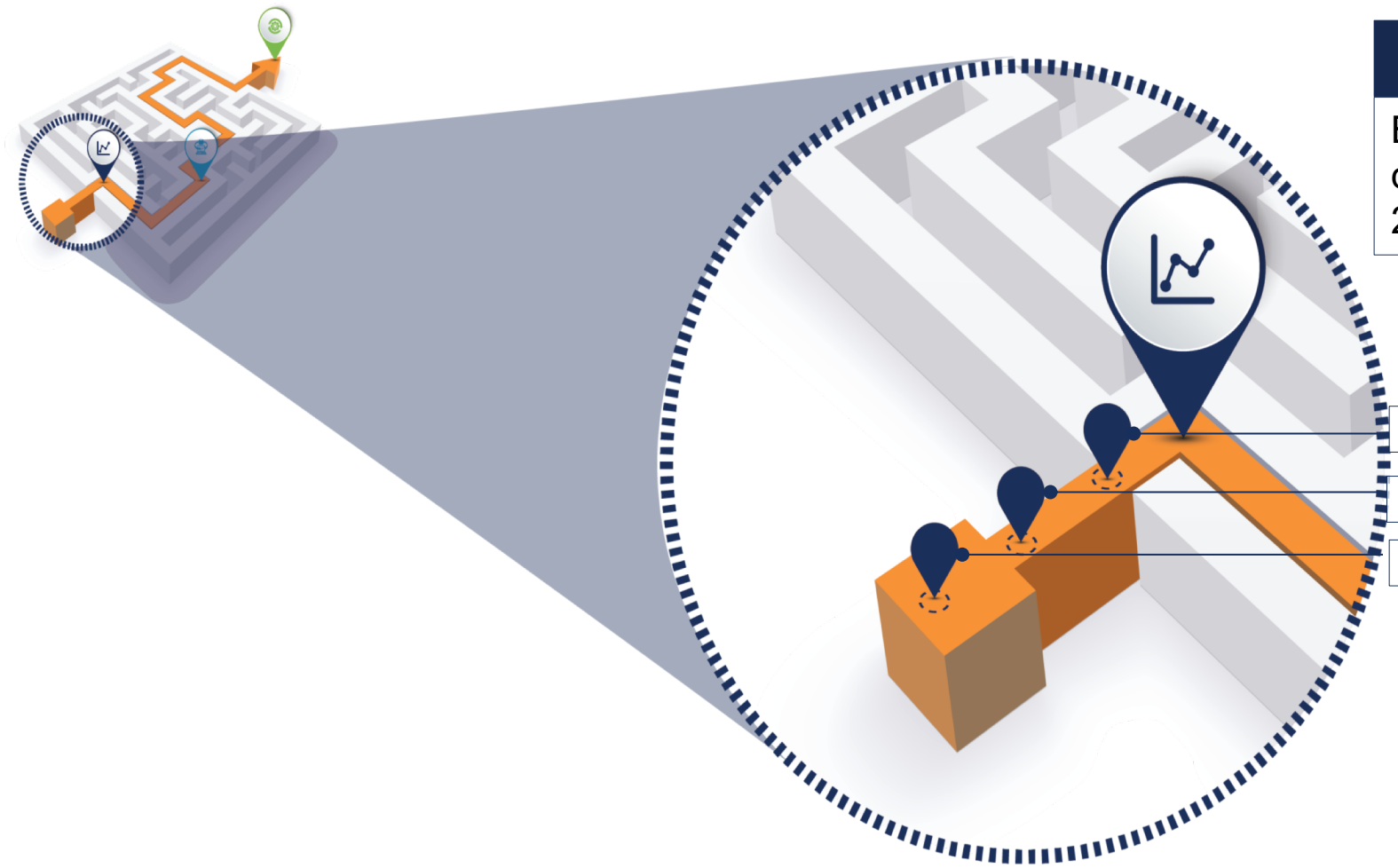
AGGREGATION

Preparation of data for FDA subject matter experts

PERFORMANCE ANALYTICS

Provide analytics and dashboards for assessing current import screening rules performance

PHASE 1: AGGREGATION

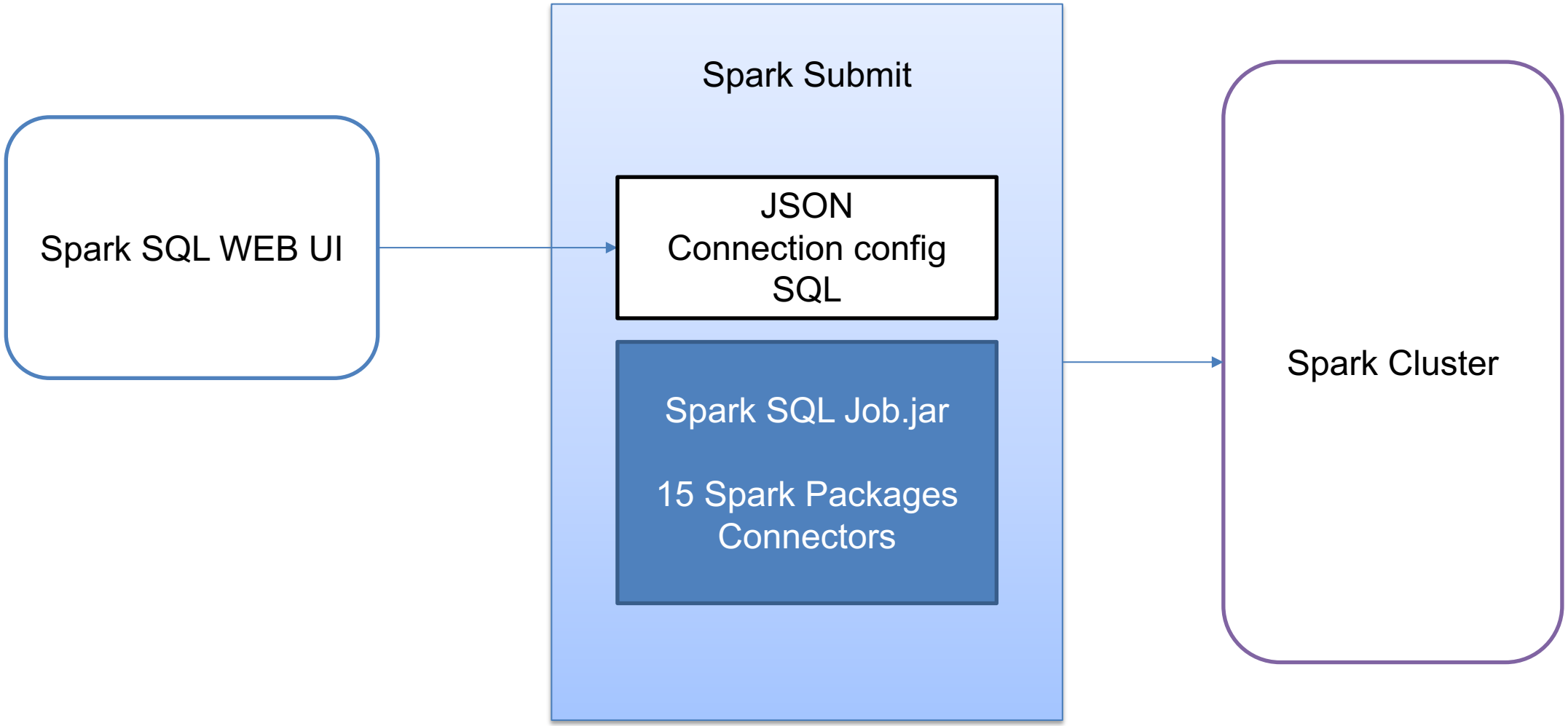


CHALLENGE

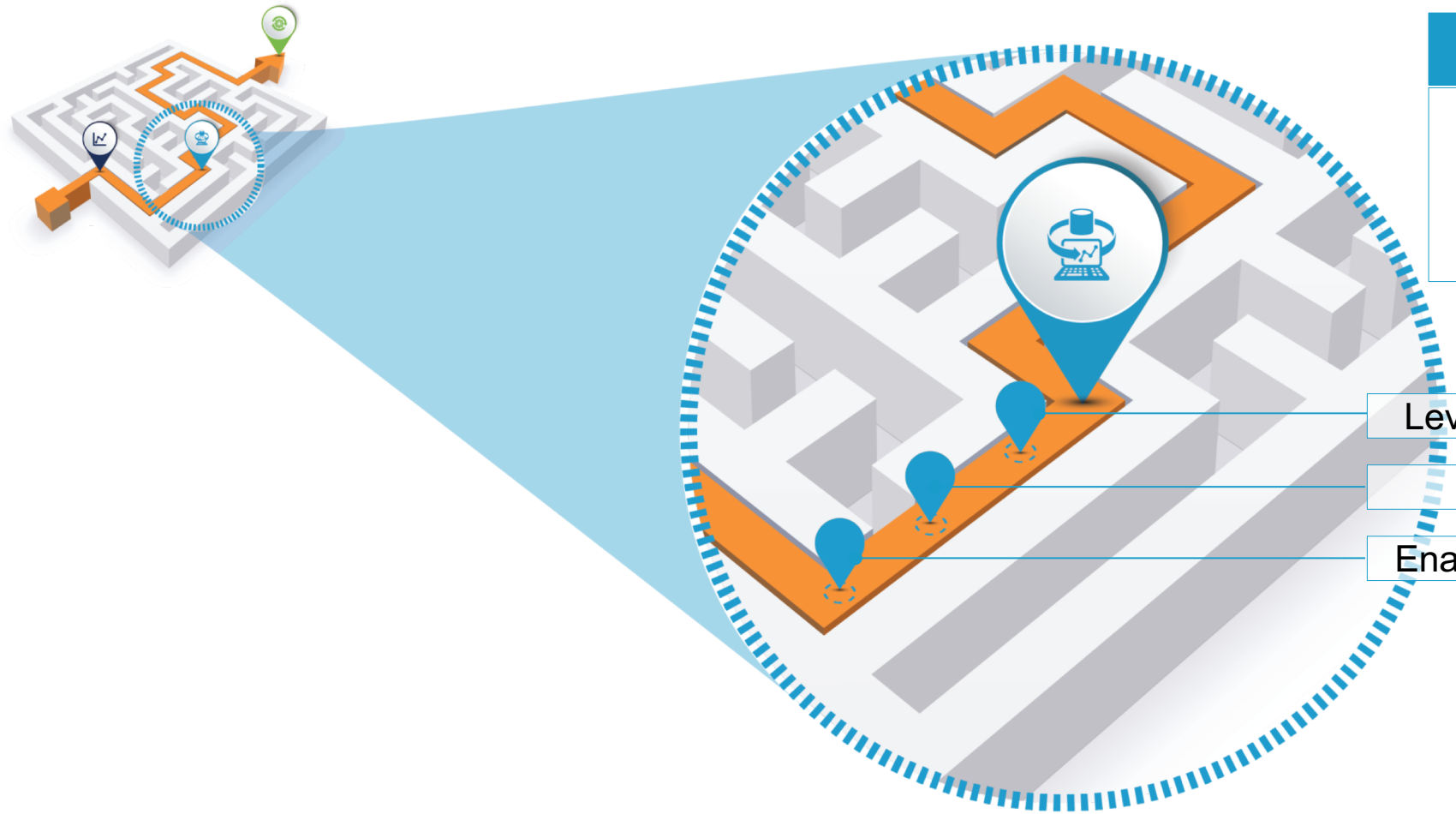
Billions of records in various database tables dating back to 2007

- Spark jobs migrate data into data lake
- Filter and join data set
- Redistribute data for performance

ODP IMPLEMENTATION AT FDA



PHASE 2: PERFORMANCE ANALYTICS

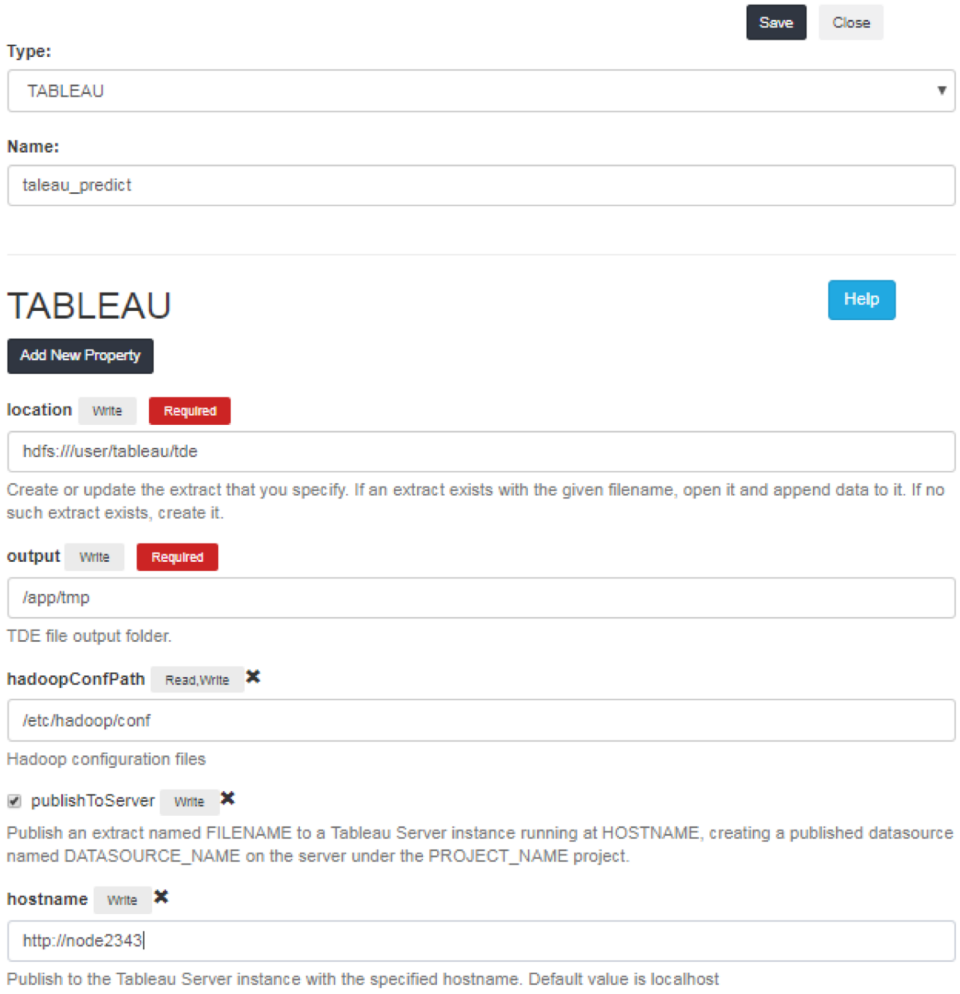


CHALLENGE

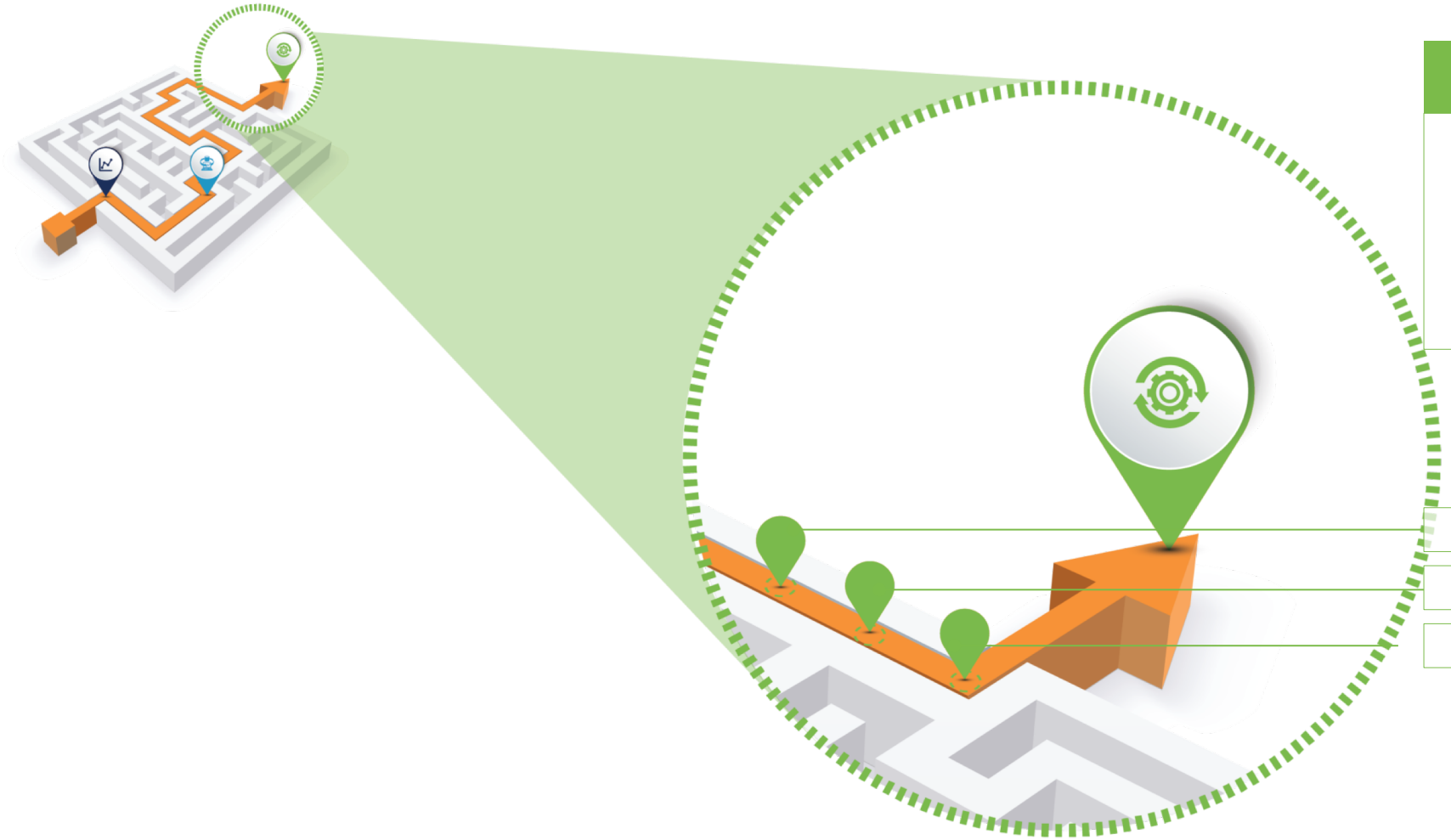
With thousands of rules in place, provide intuitive user experience and insight to provide clarity on current rules performance

- Leverage Client's Tableau Investment
- Create Tableau Dashboards
- Enable Parquet to Tableau TDE format

TABLEAU CONNECTOR



PHASE 3: CURRENT PHASE - BUSINESS OPTIMIZATION



GOALS

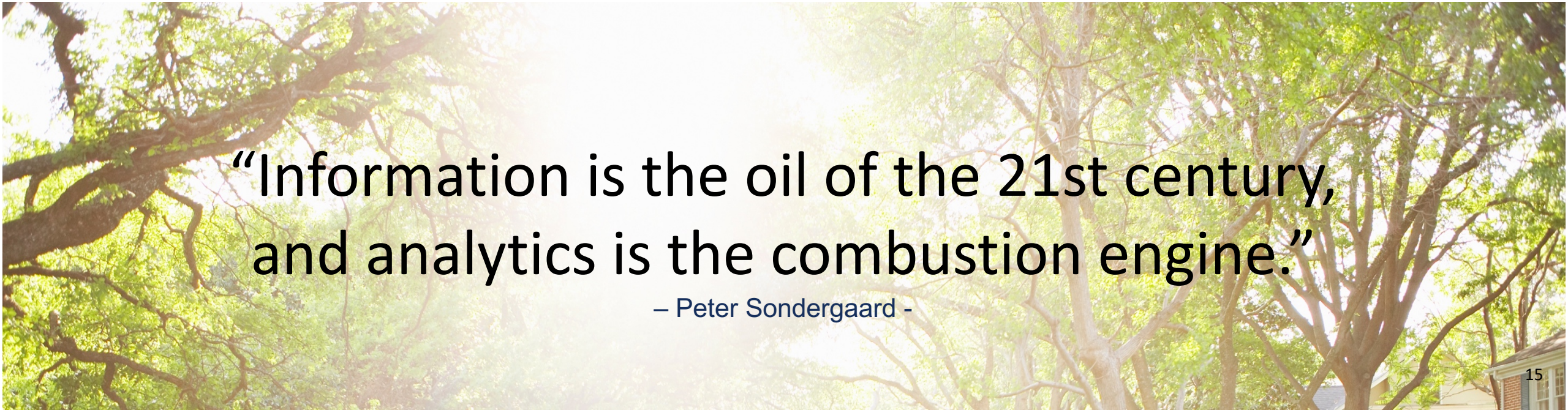
Automate use of data captured during imports entry review and other data to ascertain risk of products to improve entry review process

- Recommend changes to model
- Determine projected performance
- Determine impact to organization

DEMONSTRATION OF ODP

WHAT WE LEARNED

- The use of Spark further facilitates innovation due to its ability to quickly experiment and demonstrate business value
- Due to regulatory mandates, there will always be a human factor in the final decision



For more information about Open Data Platform, contact:

Jonathan Chu

chu_jonathan@bah.com

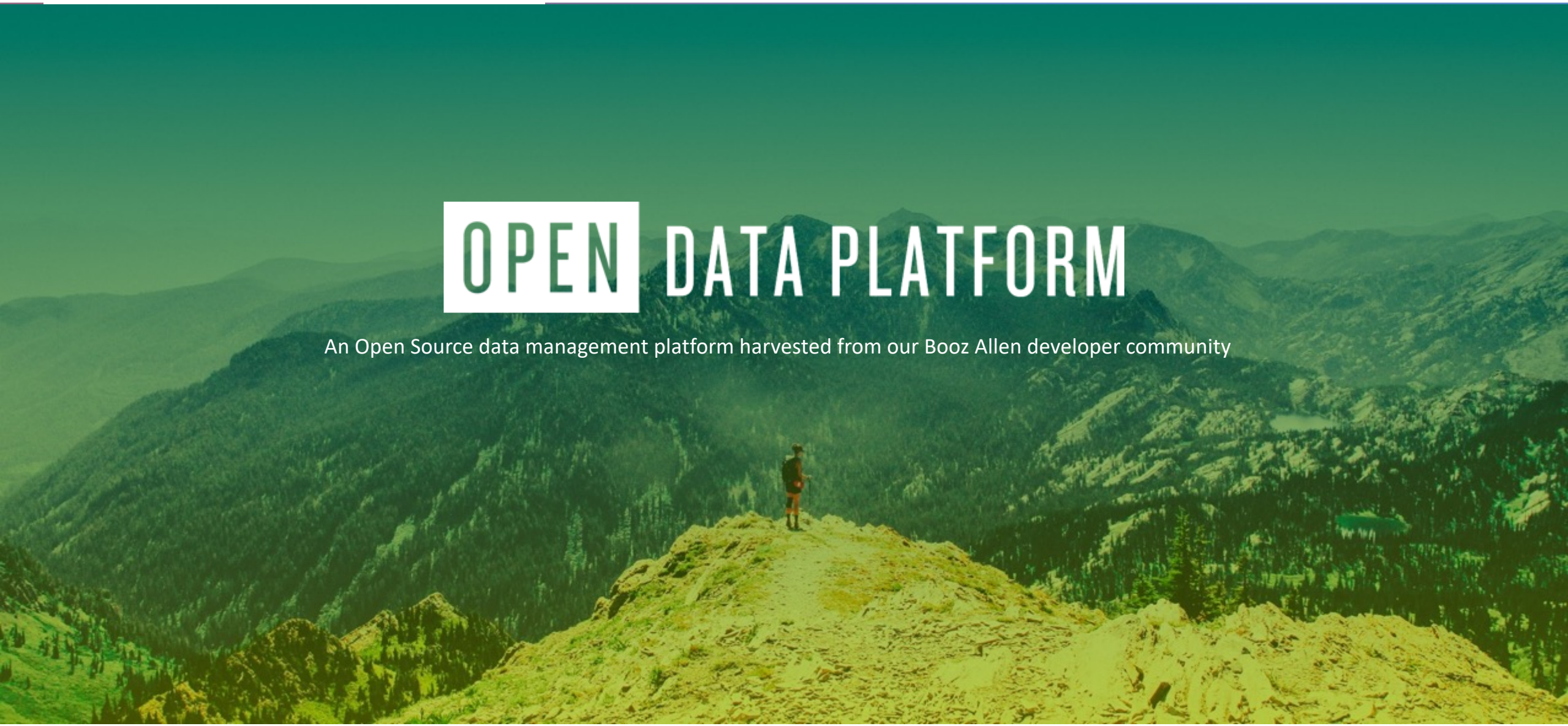
Kun Ei Kang

kang_kun@bah.com

opendataplatform@bah.com

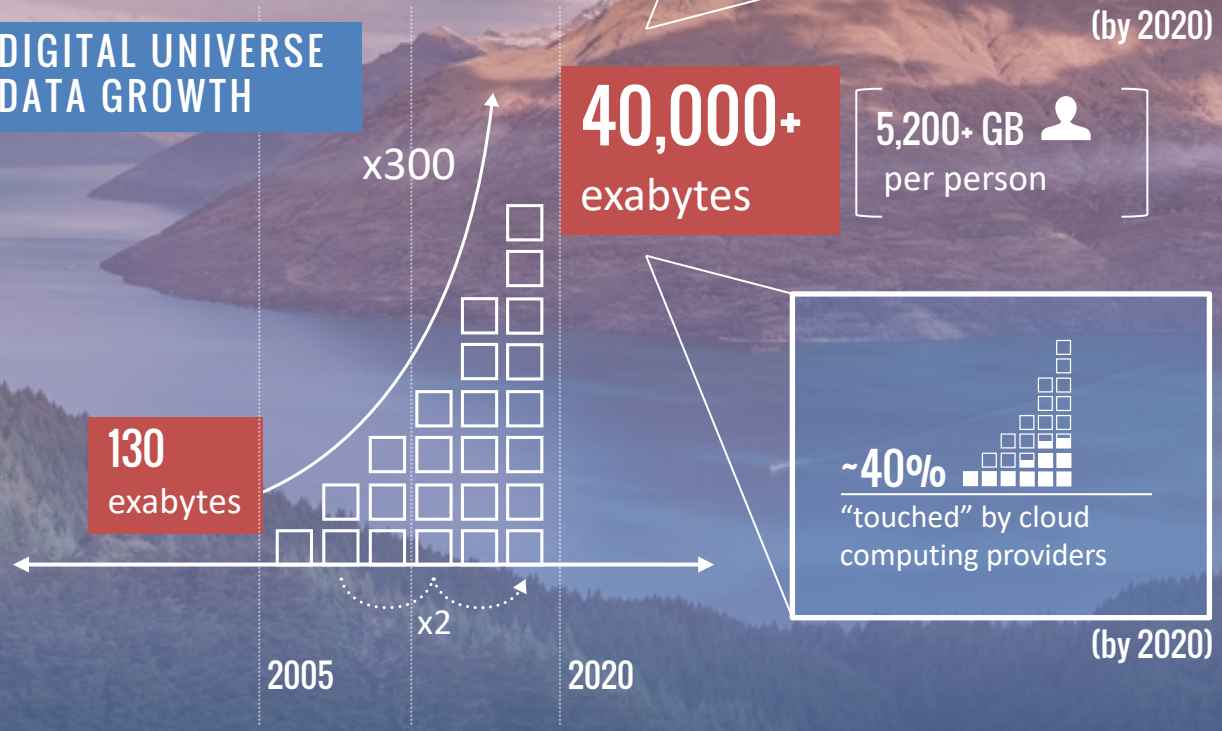
https://boozallen.github.io/opendataplatform/

ODP REFERENCE SLIDES


OPEN DATA PLATFORM
An Open Source data management platform harvested from our Booz Allen developer community

DATA CHALLENGES CONTINUE TO GROW EXPONENTIALLY

DIGITAL UNIVERSE DATA GROWTH



40,000+ exabytes

5,200+ GB per person

~33% of the digital universe will be valuable if analyzed (by 2020)

~40% "touched" by cloud computing providers (by 2020)

GROWTH OF SENSITIVE DATA REQUIRING PROTECTION

of which < 50% is protected

unprotected data will grow by a factor of 26

GROWING NEED OF DATA GOVERNANCE

Gartner estimates data lakes will still not have effective metadata management until at least 2018

By 2020, information governance and policies will be managed by metadata alone

Poor quality of data costs an average organization $13.5 million per year

*Forrester estimates that companies are only making use of **12%** of the data they own*

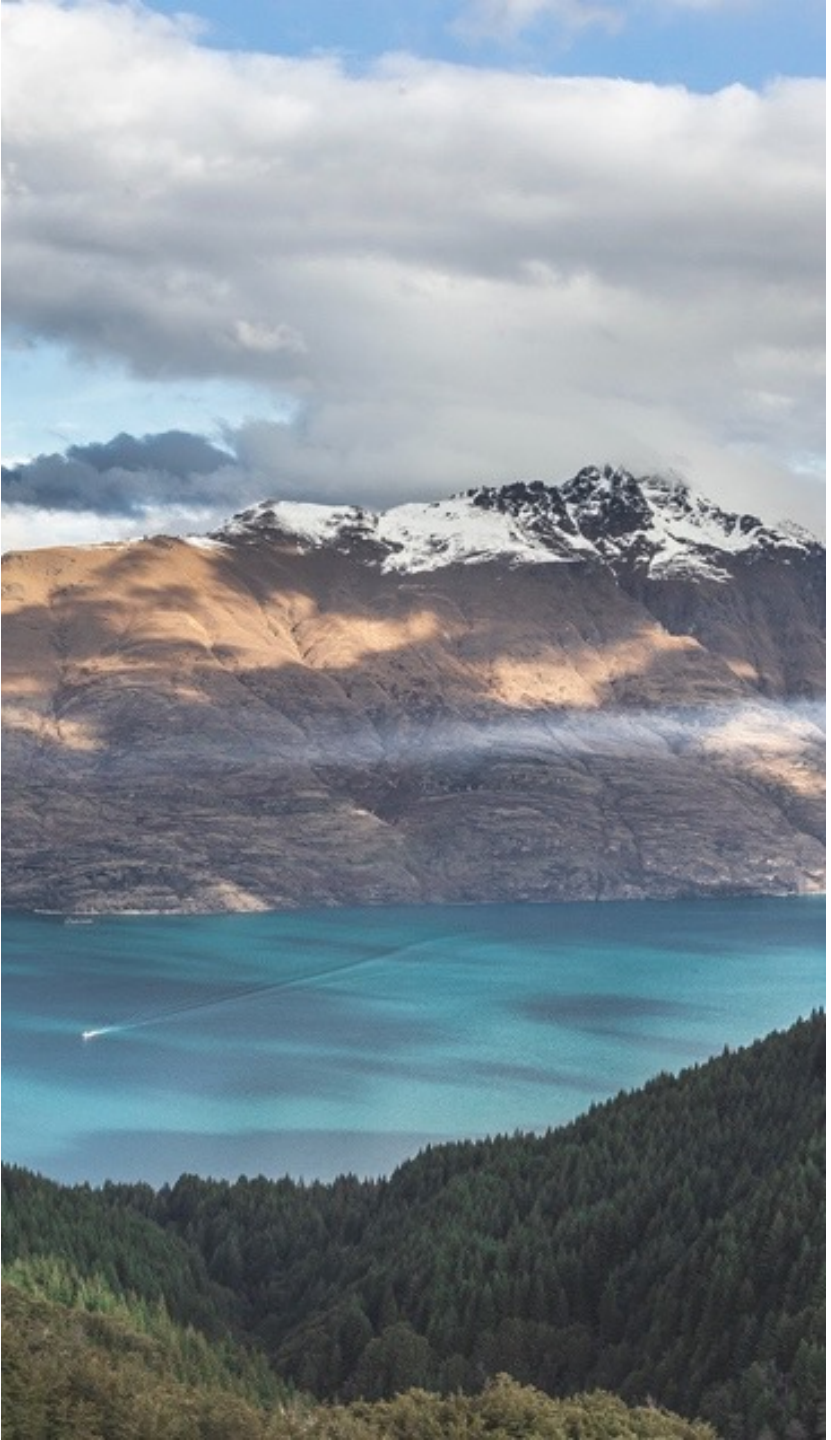
SOURCES: IDC – The Digital Universe in 2020, The Forrester Wave: Big Data Hadoop Solutions, Gartner 2016 Metadata Management MQ, Gartner Newsroom

THE TRUTH ABOUT BIG DATA

BELIEF	vs.	REALITY
"If I combine all my data together I will have full insight into what my organization is trying to accomplish"	→	Policies, restrictions, and security need to continue to be enforced and the original origin of the data and any changes to it need to be tracked throughout the life of the data asset
"If I leverage industry technologies, I will be successful"	→	Successful organizations leverage industry technologies, but they also adopt an agile culture that enables them to try new ideas and pivot quickly based on lessons learned
"Once I have all my data in one place, I can focus on analytics"	→	Stored data needs to be cataloged to enable analysts to exploit the data. Information such as data set name, data formats, data tagging, release-ability, retention, and other metadata must be available.
"The more data I have, the more questions I can answer, the more knowledgeable my organization will be"	→	It's about the right data, combined with the right tools, for the right problem— not how much data is stored

"Through 2017, 60% of big data projects will fail to go beyond piloting and experimentation, and will be abandoned."

Gartner, http://www.gartner.com/newsroom/id/3130017



WE NEED TO EMPOWER ORGANIZATIONS TO MOVE FROM DATA CHAOS TO POWERFUL ANALYTICS THAT DRIVE INSIGHT INTO THEIR MISSION

OPEN DATA PLATFORM CAPABILITES

DATA GATEWAY
Access data across your enterprise using an open common API

DATA ANALYTICS
Derive powerful insights quickly through readily available analytical capabilities

DATA QUALITY
Refine your data either during or post-processing to add greater context to the data and enhance quality

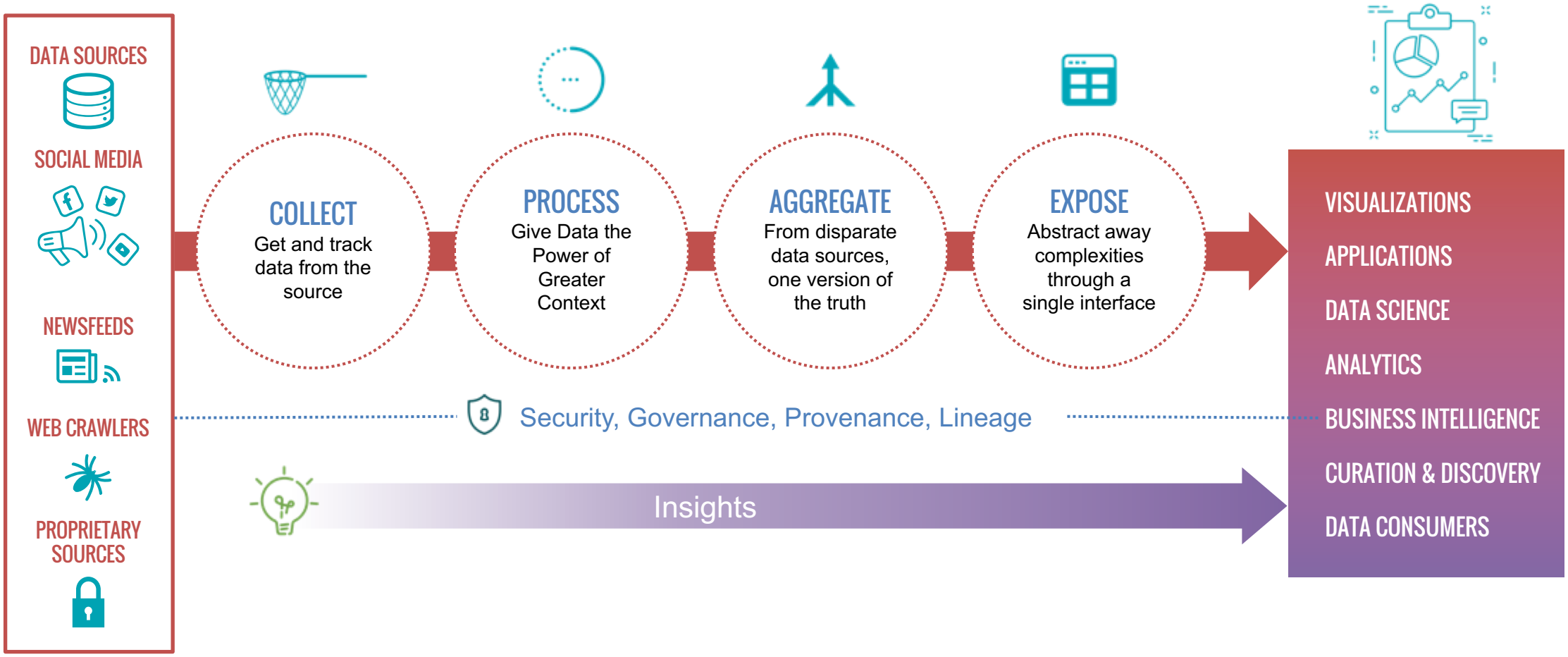
DATA SECURITY
Ensure security is applied and data is protected from the point it enters to the platform until it leaves

DATA INGEST
Fully established and orchestrated Docker-based ingest pipelines supporting RDBMS, Web Services, JSON, HL7, ICPUBS, and more

DATA GOVERNANCE
Gain insight into the health of your data feeds, users of your data (both human and system), metadata and policies, and the lineage of data across the entire platform

PLATFORM AUTOMATION
Fully automated deployment ODP enabling repeatable deployments with no manual intervention

ODP TURNS DISPARATE DATA INTO ONE VERSION OF THE TRUTH



EXCHANGE – FIND & CURATE

The power to take control of your data

INTEGRATED TEAMS & APPLICATIONS: Dissolve technical and cultural barriers by building a community for all skill levels. ODP empowers both technical staff and non-technical staff to get value from data.

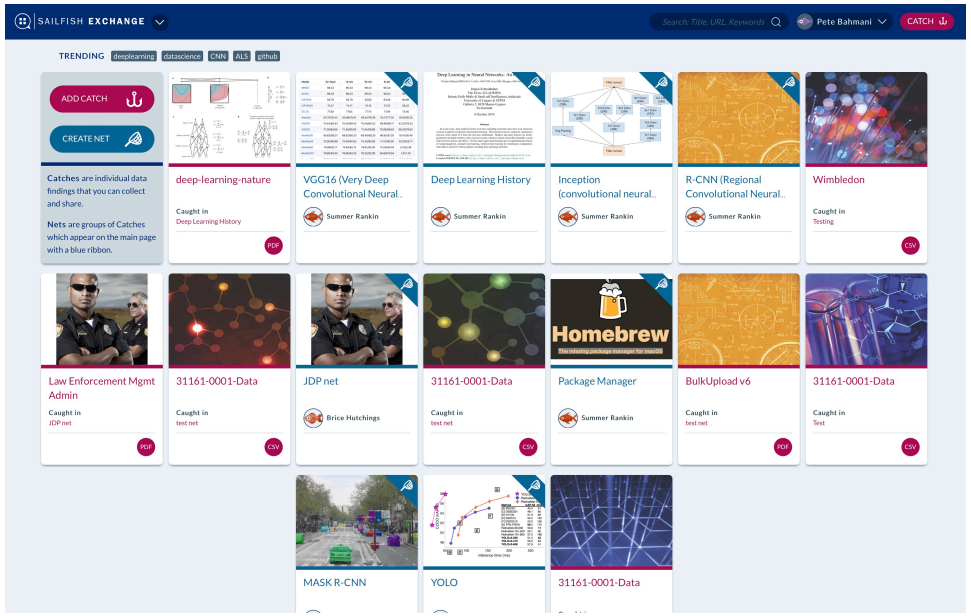
FILE TYPE AGNOSTIC: Links directly to any URL or data file, anywhere, using our intuitive Web extension.

CUSTOMIZED CURATION: Organizes data sets on clients' terms. Collects data sources into custom "nets," then curates them by topic.

DATA-GATHERING COMMUNITY: Users can create a profile, curate data, and gain followers. A social data community levels out the playing field to find quality data.

CROWD-SOURCED QUALITY: The Comment, Like, and Star Rate features give everyone a say about the quality of data sources.

CUSTOM SOLUTION PACKAGES: One size does not fit all. Every organization faces unique challenges, and no one should settle for a standard, shrink-wrapped, stand-alone tool.



EXPLORE – ANALYZE & DISCOVER

Data science & machine learning for the masses

EMBRACE MODERN ANALYTICS: Gain greater data insights through machine learning, natural language processing, advanced data querying, data curation bookmarking.

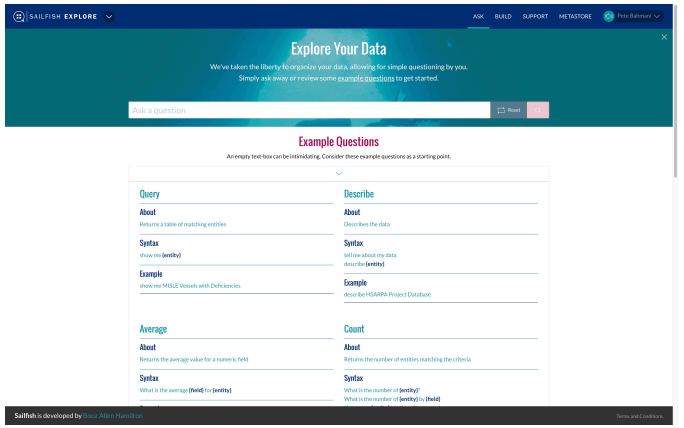
GROW A DATA SCIENCE CAPABILITY: More than just tools, ODP helps build a data science community and nurtures the growth of an analytic capability.

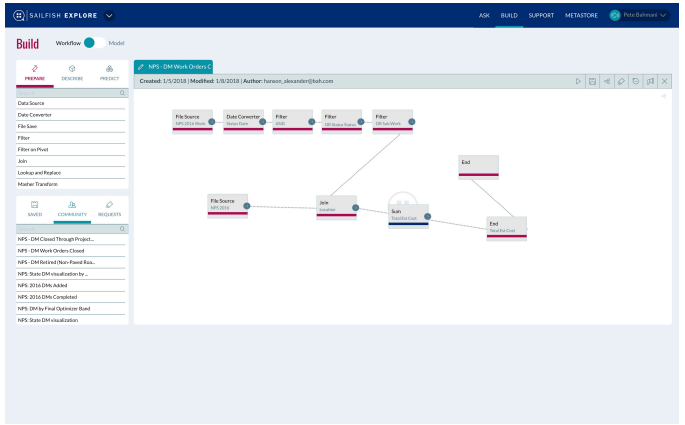
API TO ANALYTICS: Downloads data locally with one click, or sends data via API to Sailfish Explore.

NATURAL LANGUAGE INTERFACE: Users can ask "plain-English" questions about the data—no special query language is required.

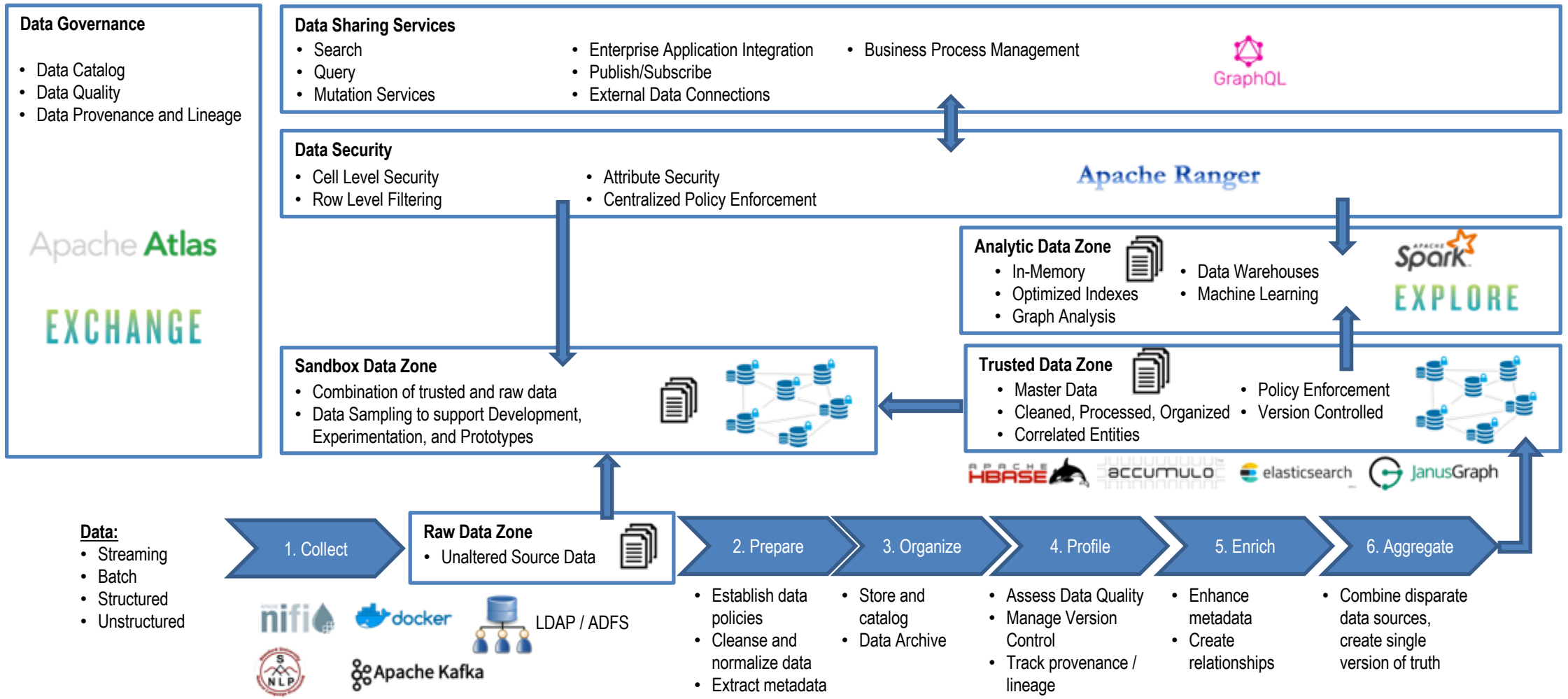
VISUAL QUERY BUILDER: Users can simply drag-and-drop to build complex queries, without coding.

WORKFLOW MANAGEMENT: Users can think through their analytical approach in real time, saving their workflow, sharing it with others, and/or scheduling it ahead of time.





ODP ARCHITECTURE AND TECHNOLOGY STACK



TYPICAL IMPLEMENTATION STRATEGY

Proof of Concept

2-3 Weeks

- Demonstrate technical capabilities in Booz Allen environment
- Target single contextual use case
- Public data sources

Ex. For a prospective commercial client, we ingested publicly available geospatial, health, and financial information in 3 weeks to showcase our ability to respond to health events worldwide

Pilot

6-8 Weeks

- Demonstrate technical capabilities in Client sandbox environment
- Sample client data sets ingested
- Pilot use case implemented against client needs

Ex. For a government client, we ingested multiple acquisition data sets to produce example analytics and enable the client to reduce license costs while increasing visibility across functions of the enterprise

Operational

Varies by Customer

- Operational deployment in Client environment against full client needs
- Planning, Installation
- Accreditation
- Client data sets ingested
- Monitoring/management of Open Data Platform
- Platform optimization based on operational use
- Upgrades

Ex. For a government client, we support a multi-year deployment ingesting over 100 feeds to support strategic analysis of threats against US assets and provide Defense/Intel level data security

WHY LEVERAGE THE ODP?

- We have packaged our experience and expertise into an automated package to give organizations a “hot-start” to analyze data quickly
- We have researched, integrated, and hardened best-of-breed Open Source Software using our experience gained in Defense, Intel, Civil, and Commercial deployments
- It is backed by a community of Big Data Professionals, Certified Cloud Architects, and Security Professionals to ensure successful, low-risk delivery
- It solves the challenges with deploying and managing a platform, wrangling and managing data feeds, and integrating best-of-breed Open Source software

ODP enables Booz Allen to partner w/ clients in an Open Architecture

A GLIMPSE OF OUR BIG DATA EXPERIENCE

Plus Commercial Pharmaceutical, Gas & Oil, and Telecommunications

* Defense / Intel Deployments on SIPRNET and JWICS