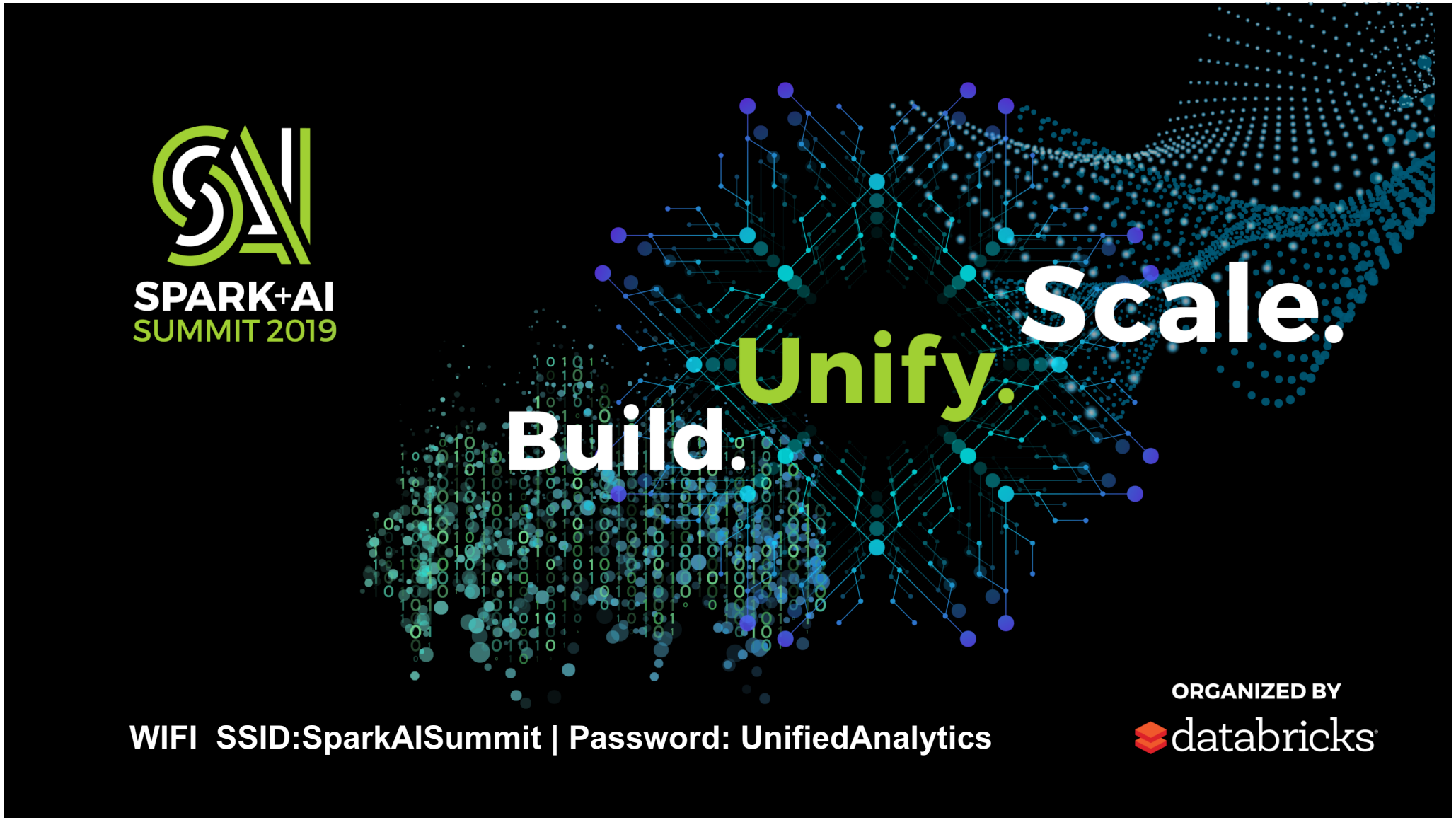

SPARK+AI
SUMMIT 2019
Build.
Unify.
Scale.
WIFI SSID:SparkAISummit | Password: UnifiedAnalytics
ORGANIZED BY
databricks

How Australia's National Health Services Directory Improved Data Quality, Reliability, and Integrity with Databricks Delta and Structured Streaming

Peter James, Healthdirect Australia

#UnifiedAnalytics #SparkAISummit

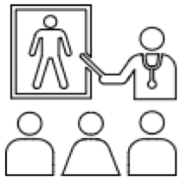
About Us

Healthdirect Australia *designs and delivers innovative services for governments to provide every Australian with 24/7 access to the trusted information and advice they need to manage their own health and health-related issues.*

Our Impact

WEB & TELEHEALTH

3+ Million Helped
Each Month



Website ranked 1st in
Health Literacy

NATIONAL HEALTH SERVICES DIRECTORY

Accessed
20+ times / second

~500K
Transactions per day

> 1 Million
Requests per Month

80+
Hospital, Clinical, Patient
Systems

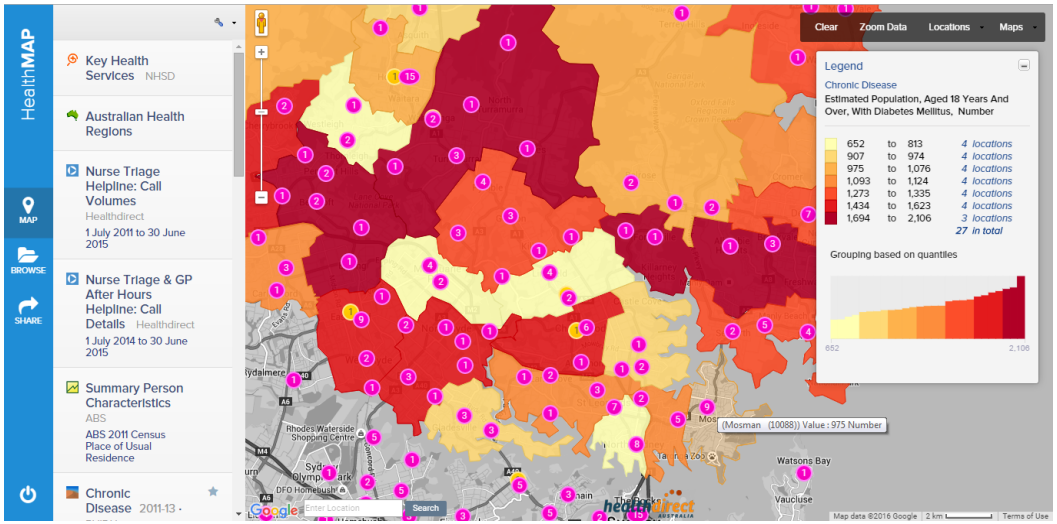
National Health Services Directory

The National Health Services Directory (NHSD) holds core information about healthcare organisations, services and practitioners.

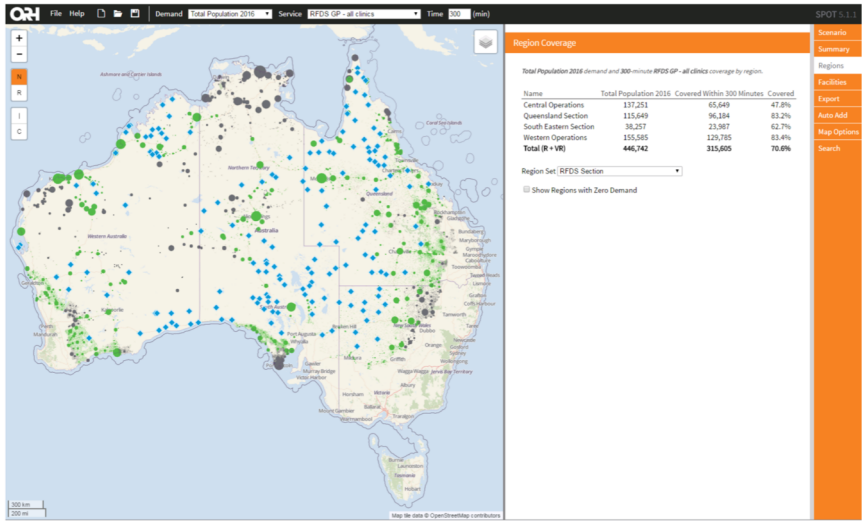
- Hospital, EMR, Registry Data sharing Arrangements
- Clinical Pathway Applications
- Telehealth and Contact Centre Services
- Support eHealth Secure Messaging Delivery
- API Integration using Fast Healthcare Interoperability Resources (FHIR)
- Consumer API's and Web Widgets
- Sector Analytics and Reporting

Supporting Health Planning

NHSD is used widely for analysis & planning



healthmap.com.au



Royal Flying Doctors Service's, SPOT platform

Landscape Changes

A Strategic Assessment defined a long-range vision of success

- Australian Health Sector promotes adoption of HL7/FHIR, Provider Directory Federation and Open Data standards
- Business drivers pushing for Cost reduction and operational efficiencies
- Architecture needed to scale to meet the objectives defined for future success
- Need to develop new Analytics capabilities in support of Sector Health Planning
- Improve Information Security to support with broader usage scenarios and regulatory changes

Challenges

Data Quality and Governance

- Availability and access to Systems Of Record
- Often no Single Source Of Truth
- High variety of data sources from Federal, State, Public/Private Hospitals, EMR, and other Commercial Vendors
- Health Domains are complex - Ontologies, Taxonomies, Thesauri, Code sets (e.g. MeSH, SNOMED-CT, ANZSIC)
- Record Linking Issues, identity harmonisation, quasi-identifiers
- Disjoint Data Governance (Network of Data Governance participants)

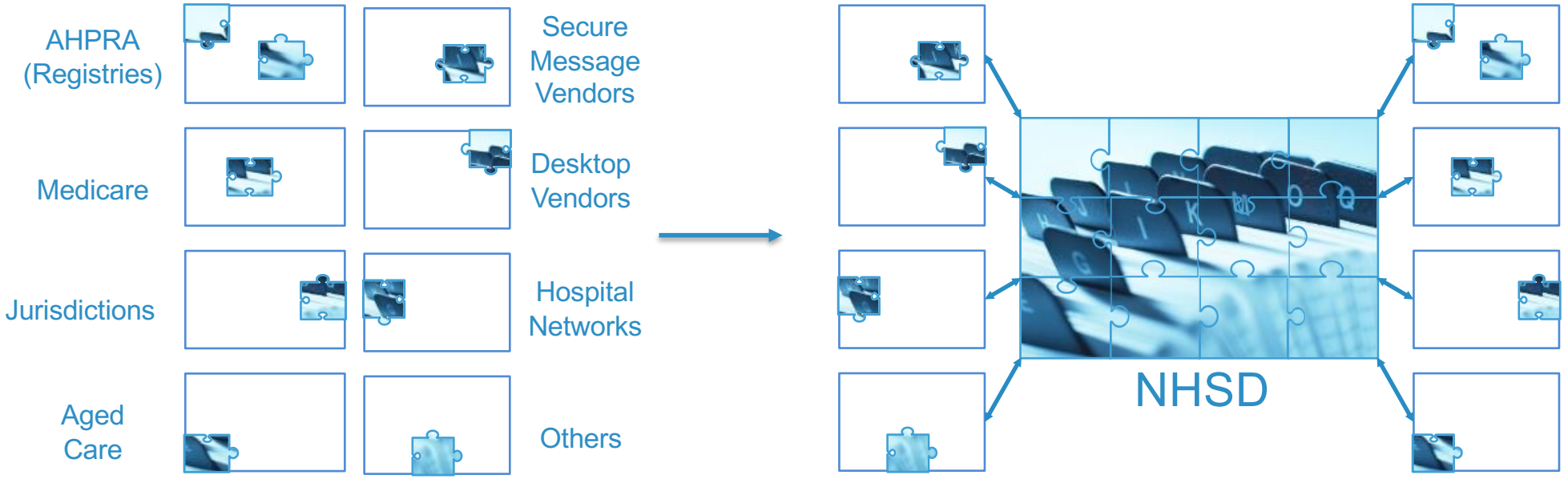
Challenges

Data Silos

- Data stored in multiple subsystems
  - File Systems, ETL and workflow transient storage systems
  - Multiple RDBMS, Multiple Schemas
  - Search Engine Indexes
- Read Inconsistency
  - Data out-of-synch between Search, Databases and Analytics storage
  - Large batch updates of low quality data leading to high error rates and process inefficiencies
  - Transactional boundaries per Data Silo no holistic end-to-end consistency
- Data Access
  - High operational overheads processes
  - Incompatible with Security boundaries

Data Silos to Data Aggregation

Single point of access to the component pieces

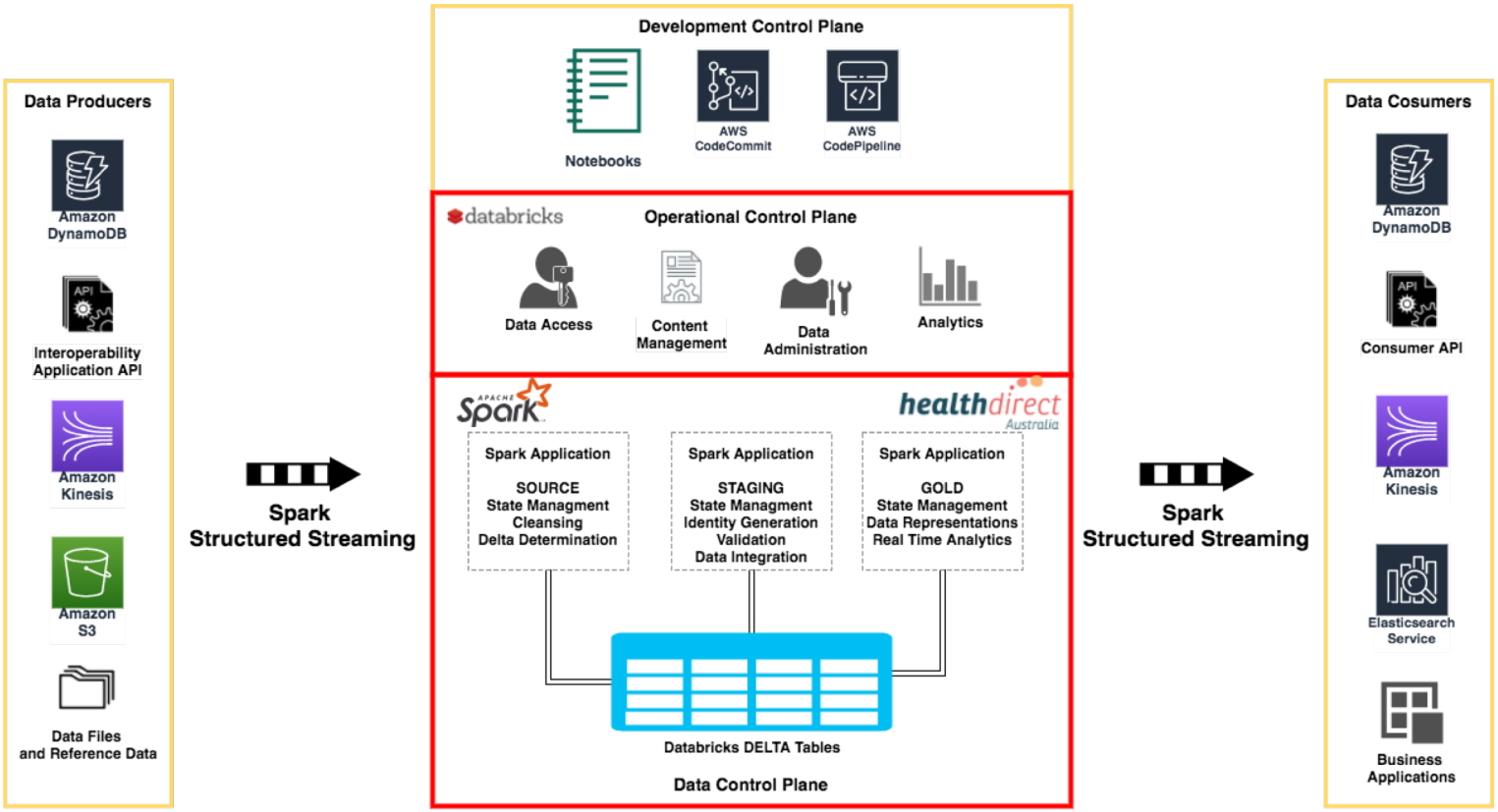


Challenges

Data Scale

- Processes need to scale to ~1 Billion Data Points
- New demands include Bookings, Appointments, Referrals, Pricing and eHealth Transaction Activity – est. 1TB p.a.
- Support Data federation and Interoperability requirements with requests growing > 58% p.a.
- Existing systems were already under pressure with batch overruns and significant administration overheads

Architecture Overview



Architecture Improvements

Databricks DELTA to create Logical Data Zones

- LANDING, RAW, STAGE, GOLD (i.e. Bronze, Silver, Gold)
- Store data ‘as-is’ either structured and/or unstructured data in DELTA Tables and Physical Partitions
- Read Consistency, Data integrity - ACID transactions
- Used DELTA Cache for frequent queries with transparent, automated cache control
- Databricks Operational Control Plane for Cluster Administration, Management functions like, Access Control, Jobs and Schedules
- Data Control Plane runs within our AWS Accounts under our Security Policy and regulatory compliance

Architecture Improvements

SPARK Structured Streaming for Continuous Applications

- Create a Streaming versions of existing batch ETL jobs.
- Move to Event based Data flows via Streaming Input, Kinesis, S3, and DELTA
- Running more frequently lead to smaller and more manageable data volumes
- Automation of release processes through Continuous Delivery and use of Databricks REST API
- Recoverability through Checkpoints and Reliability through Streaming Sinks to DELTA tables

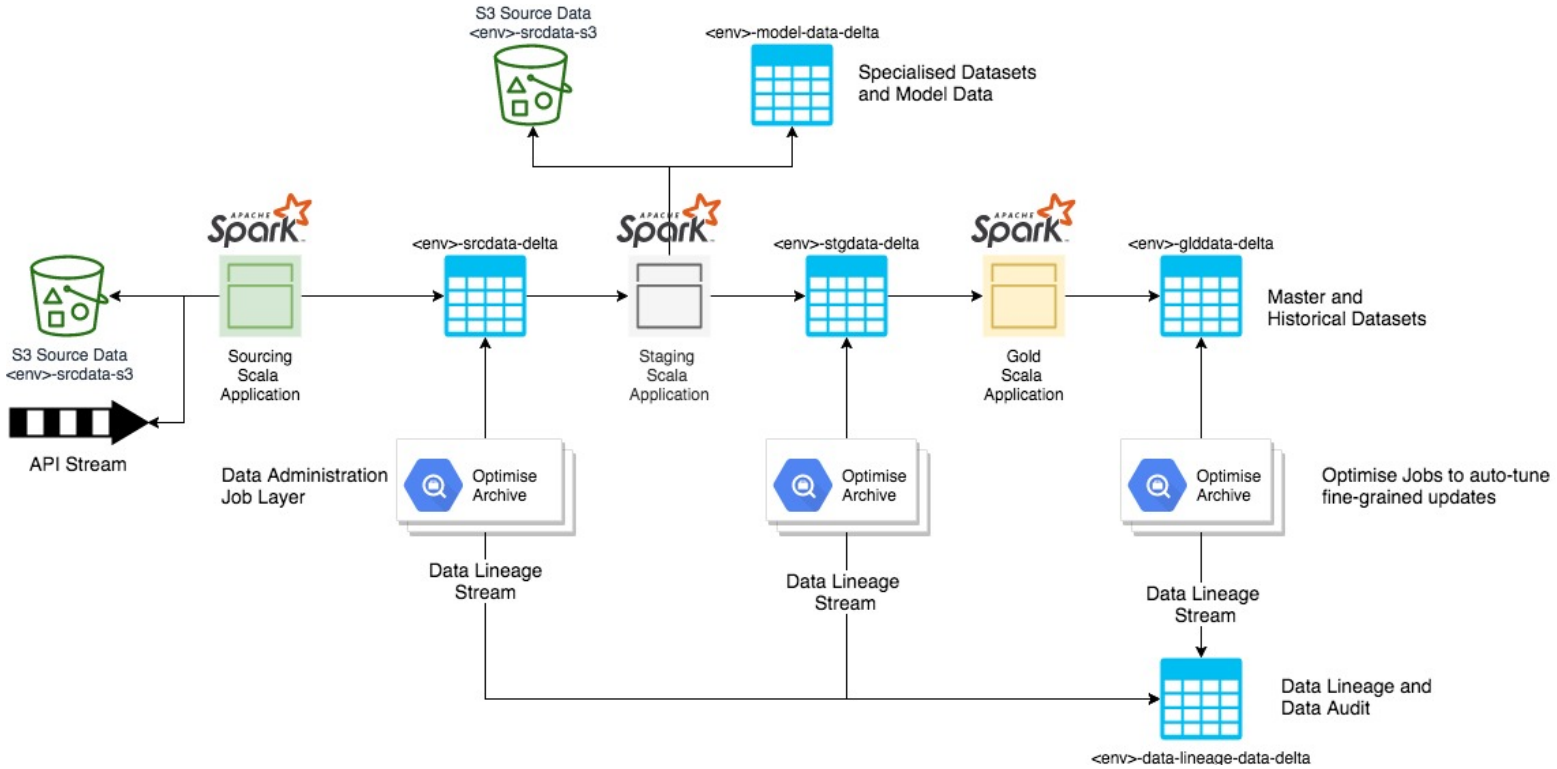
Databricks Cluster

Continuous Applications – Continued

- Keeps 'in-memory' state of object sets including comparable versions for object changes
- Stateful transformations (flatMapGroupsWithState)
- Aggregations for Data Quality measurements
- Built User-defined libraries for Data Cleansing, Validation and complex logic

```
def appendAttributeStats(
self, sourceDF,
attribute_name,
metrics = ['mean', 'min',
'max', 'median',
'num_missing',
'num_unique']):
```

Data Plane & Pipeline

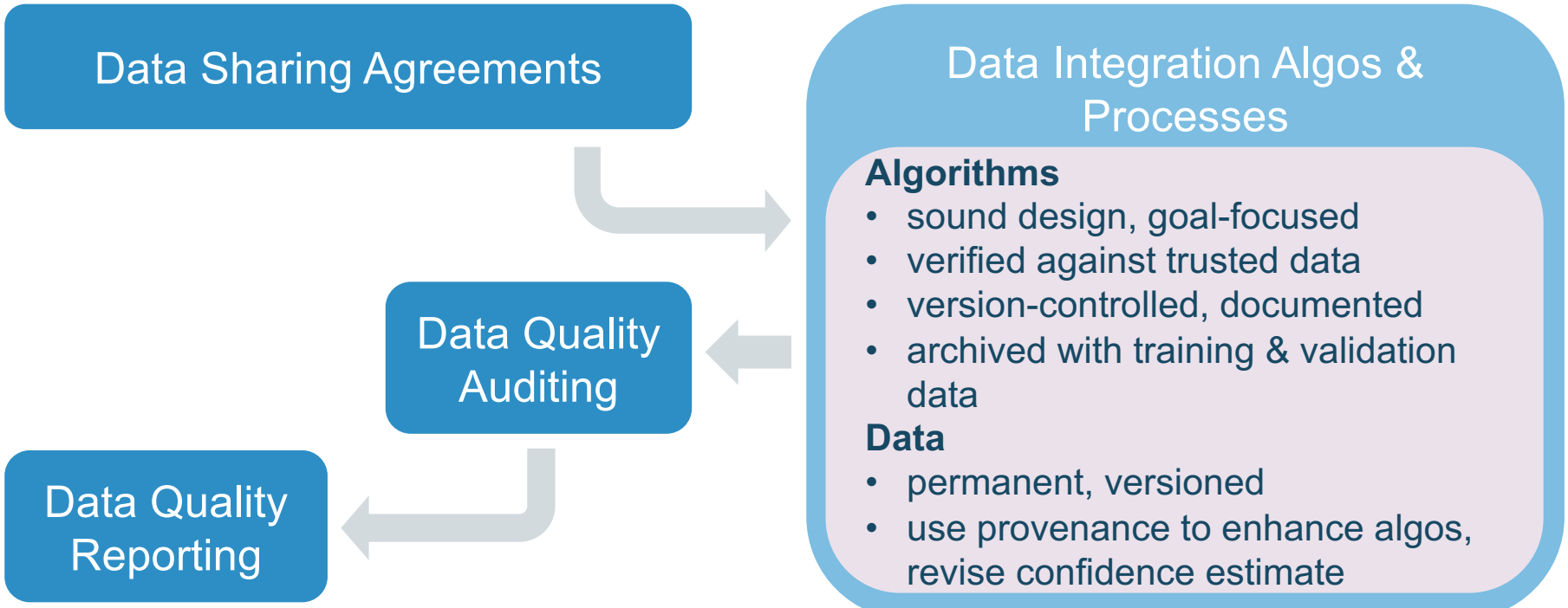


Architecture Improvements

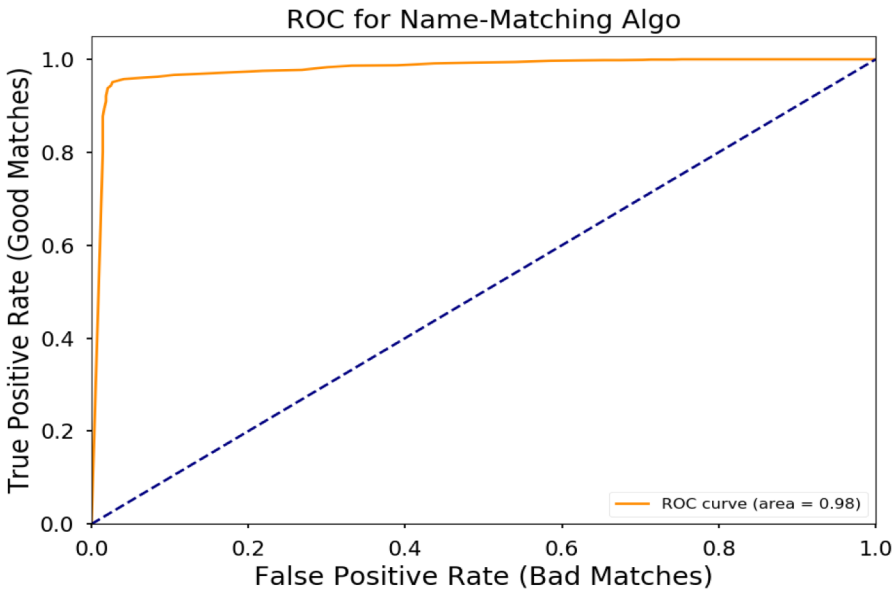
Unified Analytics and Operations

- Data cleaning & matching algorithms reliable, predictable and measurable
- Data Quality Analytics calculated at runtime, attached to object attributes, populate analysts dashboards
- Track Data Lineage and Lifecycles providing Operational visibility of where our data is at any point
- Complex processes easily broken down into simpler steps with ‘checks and balances’ prior to execution
- Significant cost reduction through decommissioning complex Administration UI in favour of notebook environment

Improved Algorithm Development & Usage



Algorithm Evaluation



ROC curve for the set-based, fuzzy name-matching algorithm

FPR	TPR	threshold
0.014433	0.800855	1.00
0.014433	0.801709	0.99
0.014433	0.831054	0.98
0.014433	0.855840	0.97
0.016495	0.895726	0.95
0.026804	0.950997	0.90

- interpretation
  - the algorithm is appropriate to the domain
  - the data is of good quality
  - un-matchable names are rare exceptions
  - a matching threshold of 0.95 gives
    TPR = 95.1%
    FPR = 2.7%

Improved Data Processing

- 95% fuzzy name matches vs. <80% and dependent on manual verification
- ~30,000 Automated Updates / mth vs. ~30,000 bi-annual & Manual
- 20 mins full load vs. > 24 hrs
- 20 million records @ ~1 million/min

Improved Data Security

- Databricks Platform provided foundational Security Accreditations
- Able to cross-walk these security compliance frameworks and localize to Australian Regulatory frameworks
- This provided significant cost/effort reduction of GRC
- Continuous Data Assurance - Security Guard-Rails monitor changes to access privileges, e.g. role elevation/conditional access, data security metadata data and auditing against data leakage and spills.

Securing it ALL

Other Benefits

- Don’t have to be big to see benefits, strategic value and a new business model
- Strategic momentum through realizing business vision with technology transformation
- Dataset Transparency and Explainable Models with well-documented lineage and quality
- Broader participation - no one holds a single parts of knowledge anymore, we extract more value in our data when everyone was able to access it
- Governance transitioned focus on operational improvements
- **Inked Agreement in August 2018 completed migration to Databricks December–Live March 2019**

SPARK+AI SUMMIT 2019

DON'T FORGET TO RATE AND REVIEW THE SESSIONS

SEARCH SPARK + AI SUMMIT

Download on the App Store

GET IT ON Google Play