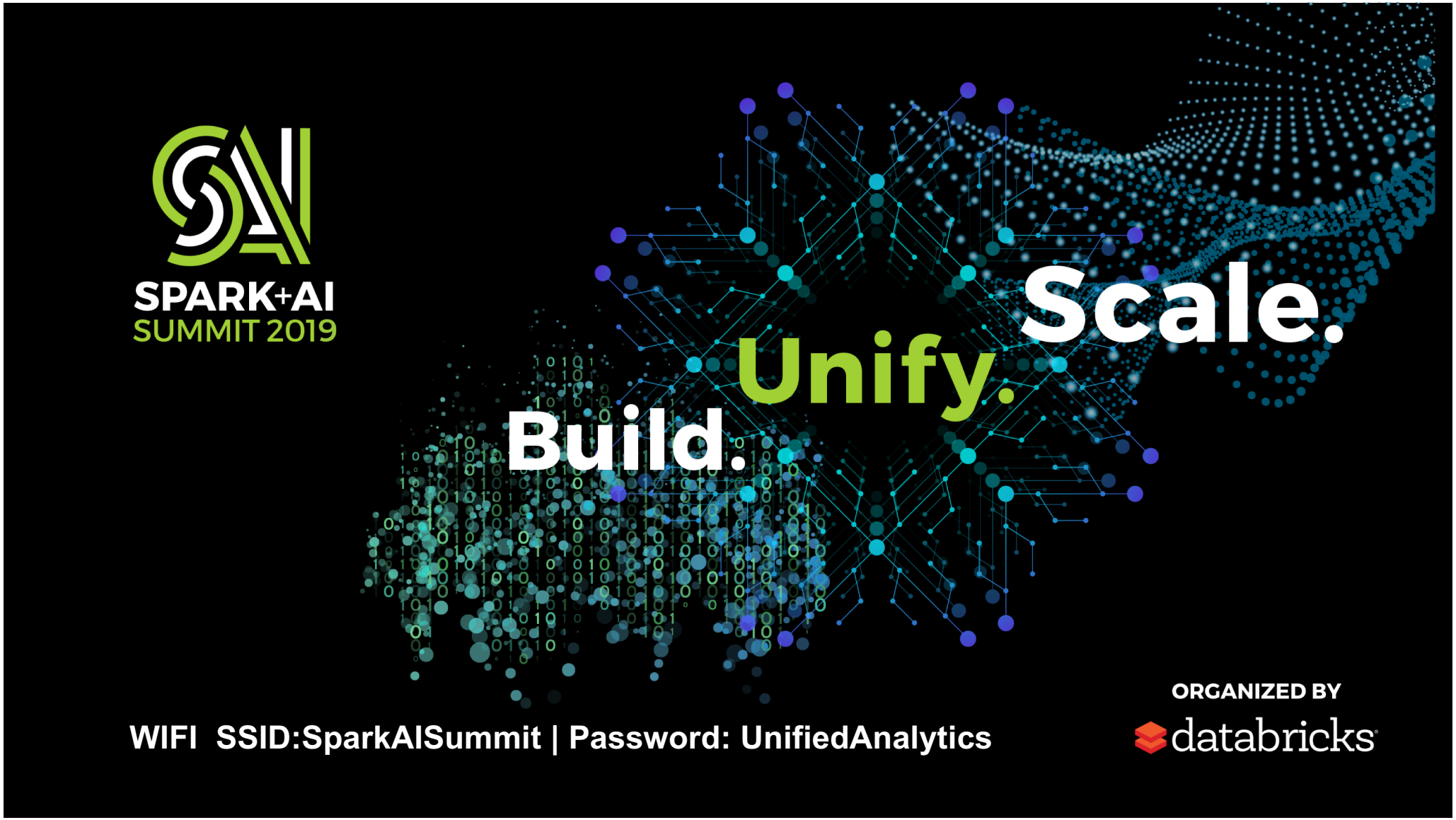

SPARK+AI
SUMMIT 2019
Build.
Unify.
Scale.
WIFI SSID:SparkAISummit | Password: UnifiedAnalytics
ORGANIZED BY
databricks

Customer Insights from 250TB+ of Data:

Lessons Learned in Data Governance and Lineage

David Newell & Geoff Oitment, McAfee

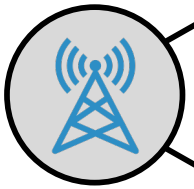
#UnifiedAnalytics #SparkAISummit

At the heart of [AI] technology is data.

— Henry Schuck, Forbes

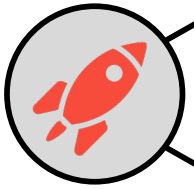
https://www.forbes.com/sites/forbestechcouncil/2018/05/02/why-data-accuracy-is-critical-to-the-evolution-of-artificial-intelligence-in-b2b-sales/#57bebedf466d

What's so hard about Big Data?



Finding Signal in the Noise

Limited Big Data Skills

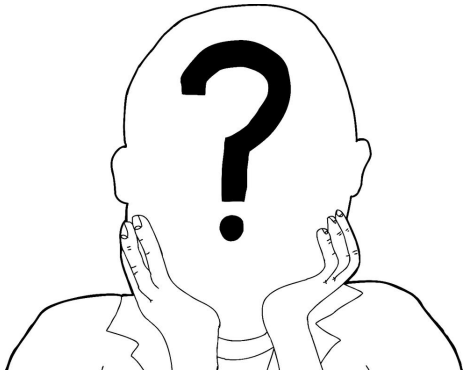


Technology Moves Fast

Siloed Data Sources

McAfee started no better

- What activity does an event track?
- Limited data science talent across business functions
- Current data systems are 15-20 years old
- Multiple data silos



Typical McAfee Data Analyst

I don't know how to start!

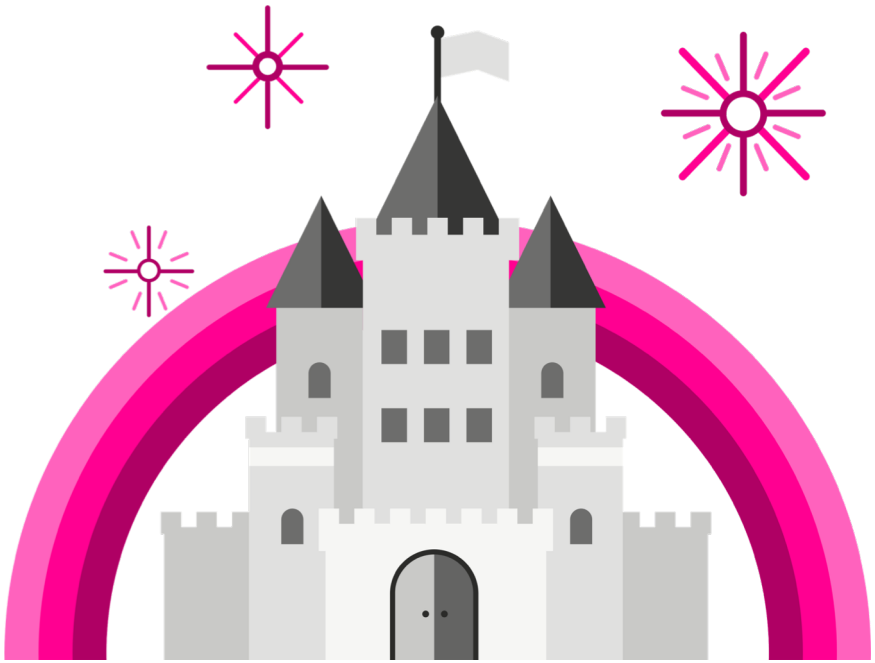
Goal: enable self-service analytics

Comprehensive Customer 360
All the data in one place

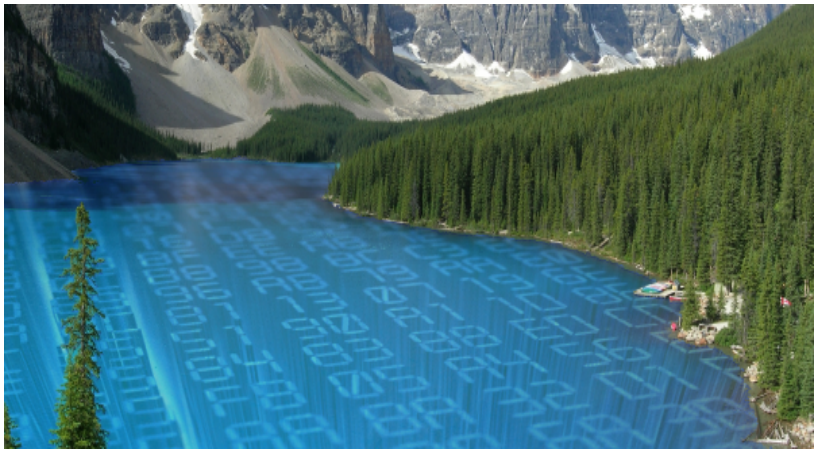
Accurate Calculations
Standardized, certified metrics

Easy to Access
Self-service in tool of choice

Scalable & Responsive
Get data fast, even in real-time



The data lake fallacy: it's magic!



vs

A data lake is not a magic wand:

- Veracity still dependent on data quality
- Volume still dependent on inputs
- Value still dependent on skill set

How we kept the data lake clean:
portal for systematic data governance

Comprehensive documentation at all ingestion points

- Data validation rules
- Source data collection driven by documentation

Centralized governance & documentation system

- Automatic documentation and governance enforcement
- Generated tracking manifests
- Data cleaned at the source

What we built:
documentation-driven pipeline

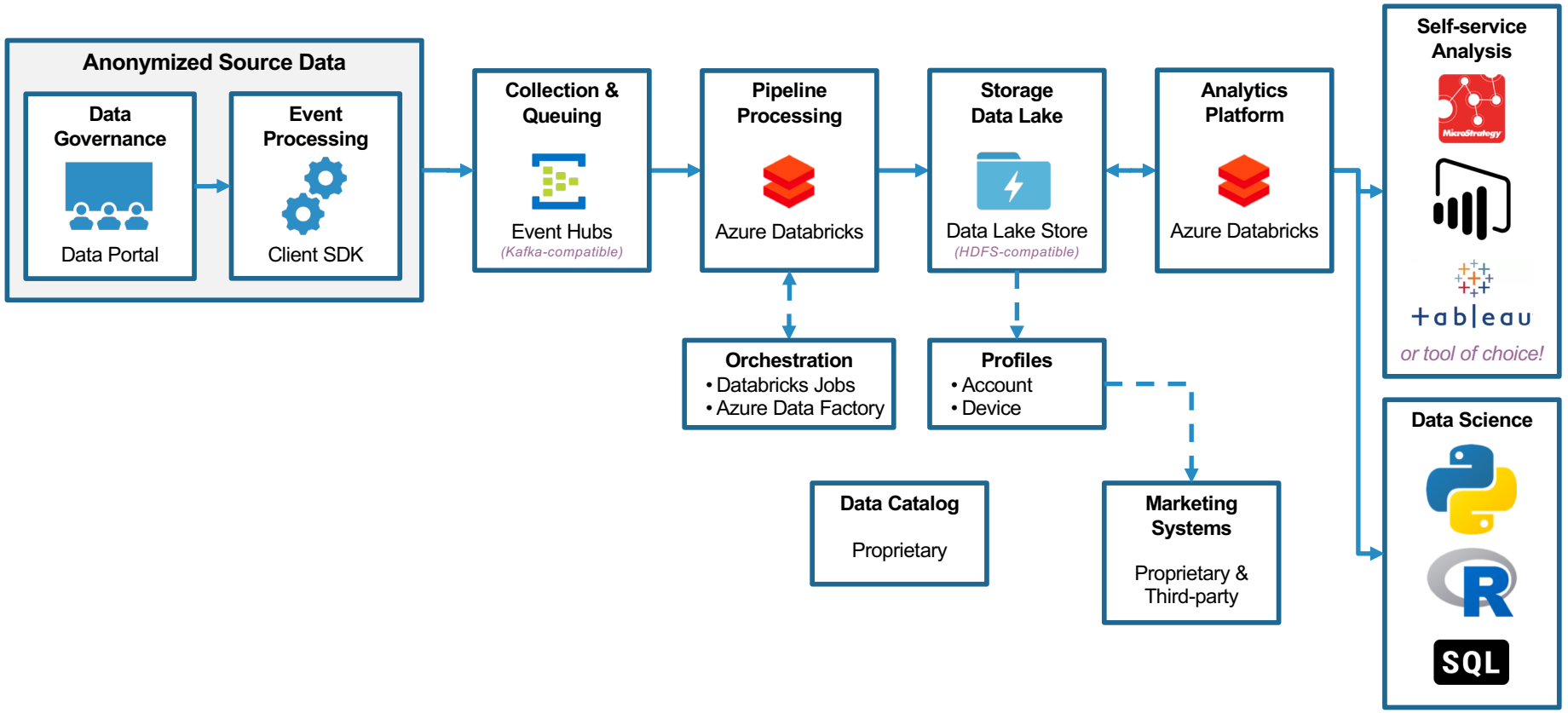
Configuration-driven design

- All ETL jobs defined in configuration
- Data lineage derived from same configuration

Integrates into the centralized portal

- Enables end-to-end data flow visualization
- Configuration generated by governance system

Architecture



Lessons Learned

Don’t compromise on data quality; you’ll regret it later

Create small, manageable features that enable quick iteration

Think ahead; don’t do something that will handicap future development

Don’t compromise on data quality; you’ll regret it later

DON'T FORGET TO RATE AND REVIEW THE SESSIONS

SEARCH SPARK + AI SUMMIT