Gartner Data & Analytics Summit
Summit 2018
22 – 23 May 2018 / São Paulo, Brazil

Gartner Summits

Hadoop and Spark: Understanding Open-Source Opportunities and Risks

Merv Adrian
@merv

Strategic Planning Assumption

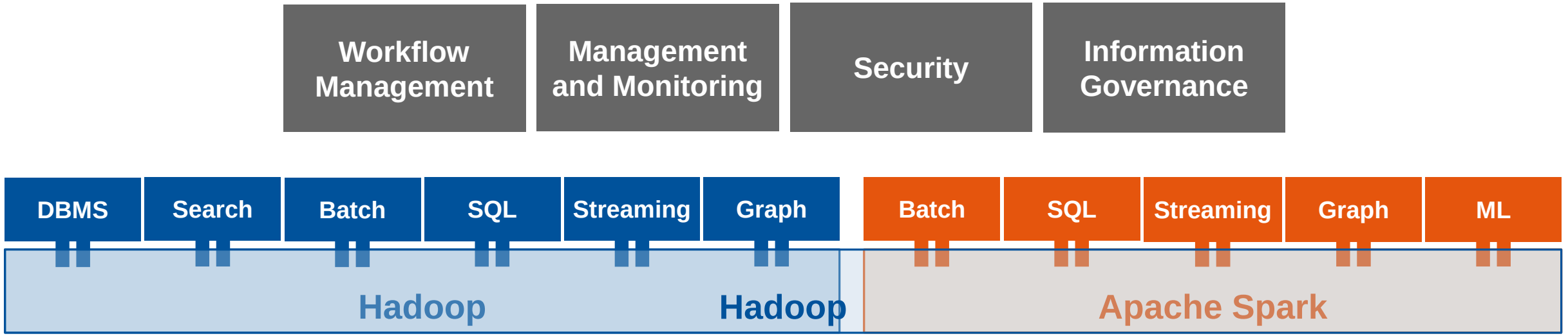
Through 2018, 70% of Hadoop deployments will fail to meet cost savings and revenue generation objectives due to skills and integration challenges.

How did we do?

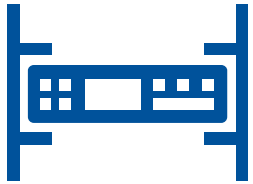
For three successive years, "in production" as a percentage of Hadoop and big data projects has been below 20% in Gartner surveys.

Gartner

Competing and Complementary Capabilities



Commodity Hardware



Appliances

IaaS and PaaS

DBMS

Containers

Gartner

Key Issues

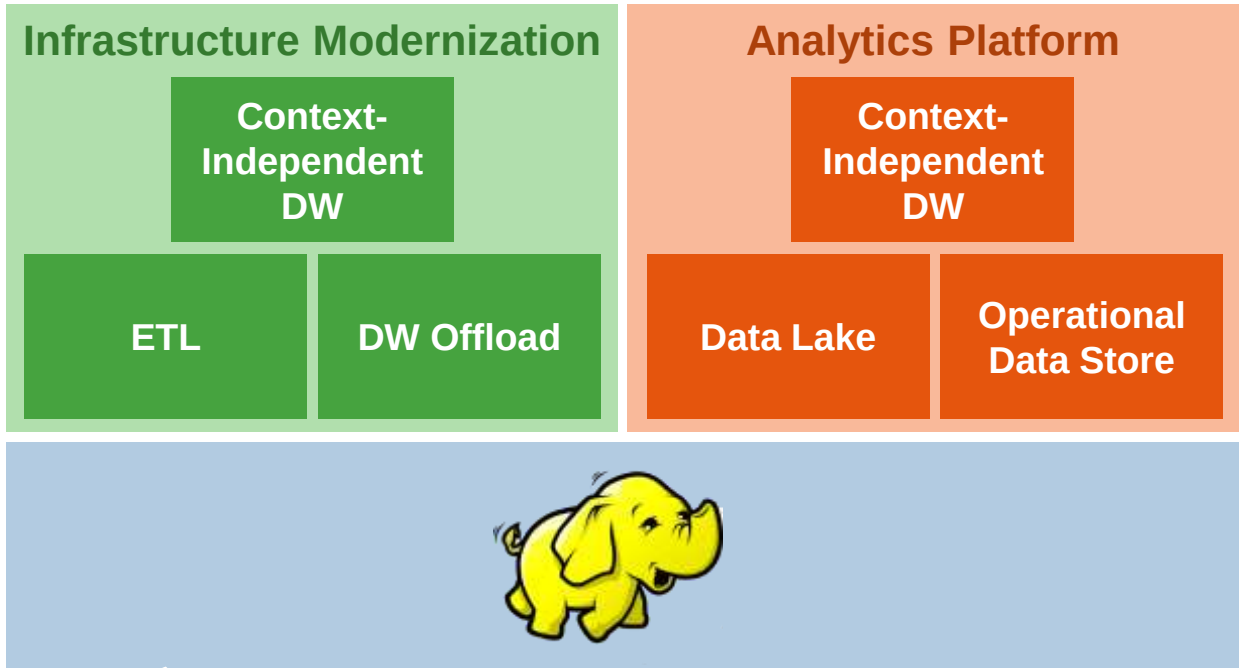
1. What are the use cases for Hadoop and Spark?
2. How will these technologies evolve over the next three to five years?
3. How do you prepare for an uncertain future for Hadoop, Spark and emerging technologies?

Key Issues

1. What are the use cases for Hadoop and Spark?
2. How will these technologies evolve over the next three to five years?
3. How do you prepare for an uncertain future for Hadoop, Spark and emerging technologies?

What Is Hadoop Used For?

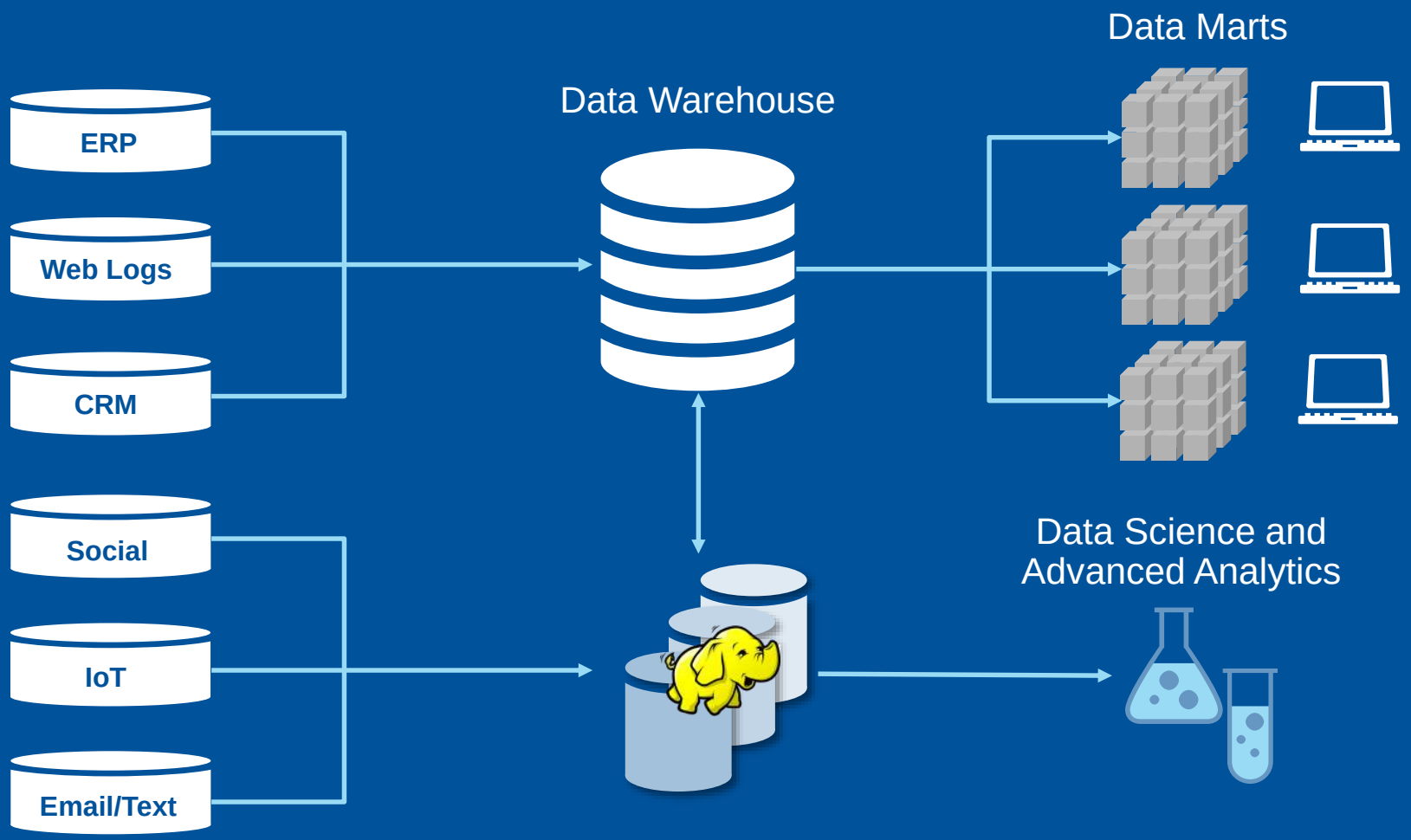
- ETL and batch processing
- Data lakes
- Data warehousing:
  - Offload
  - Contextual DW
- Operational data stores



Not all of these are equally successful in practice — as elements of the Hadoop stack improve, its capability gets closer to its ambition.

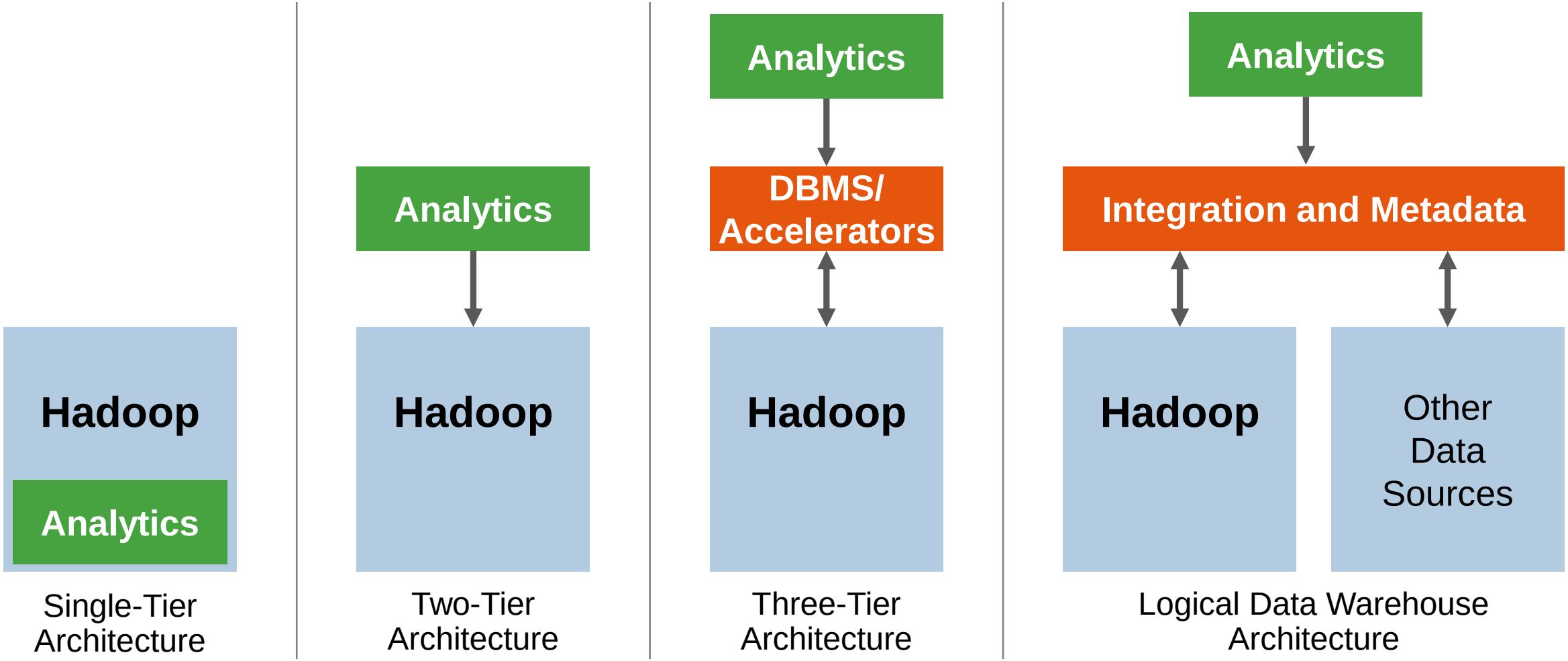
Gartner

Optimizing ETL/ELT and Advanced Analytics



Gartner

Architectural Patterns for Analytics on Hadoop



Gartner

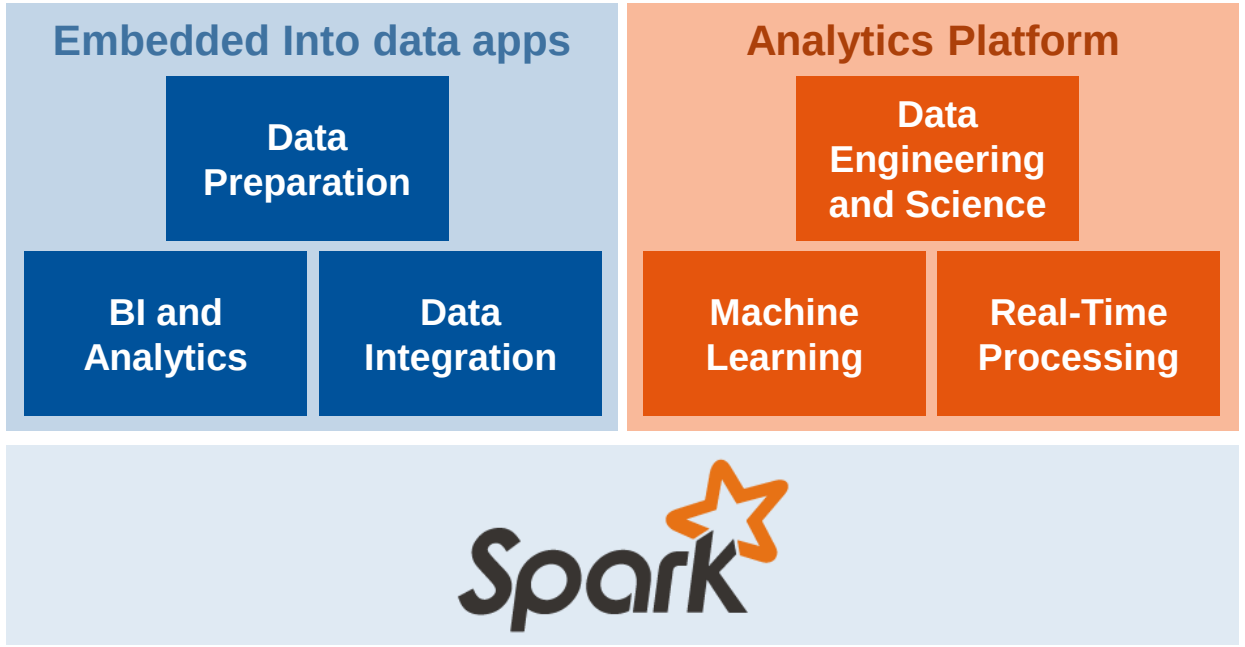
Advanced Analytics at FINRA Offloads Data Warehouse

- The Challenge:
  - SEC consolidated audit trail project pushed FINRA to rethink its analytics infrastructure
  - On-premises solution using Greenplum Database, IBM Netezza, SAS and Cloudera
  - Replicating environment in Amazon Web Services (AWS) only offered limited advantages
- Solution:
  - Worked with AWS to port HBase to run on Amazon Simple Storage Service (S3)
  - Required a shift to DevOps practices and unified teams working in parallel
- Benefits:
  - Saved $1 million annually vs. running in a separate Hadoop instance

Gartner

What Is Apache Spark Used For?

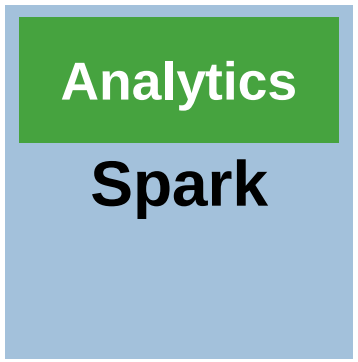
- Data integration/Log processing
- Internet of things
- Business intelligence
- Advanced analytics and ML:
  - Fraud detection
  - Recommendation systems



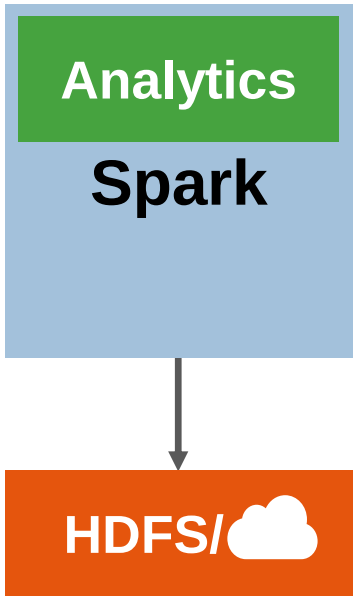
Spark is in every Hadoop distribution — but it is often used "without Hadoop." It is often provided by other vendors, including Databricks.

Gartner

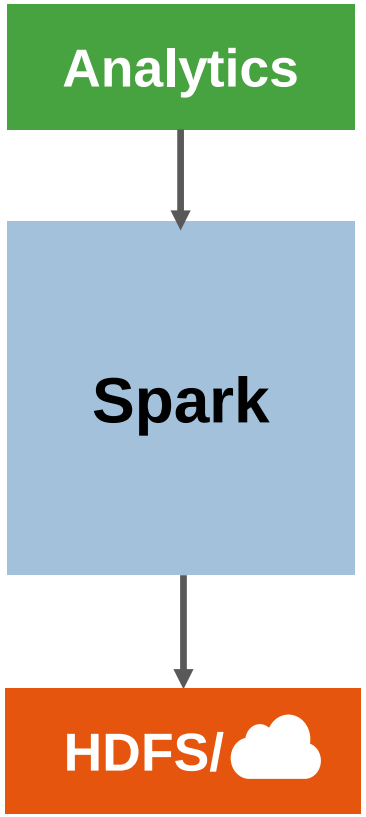
Architectural Patterns for Analytics With Spark



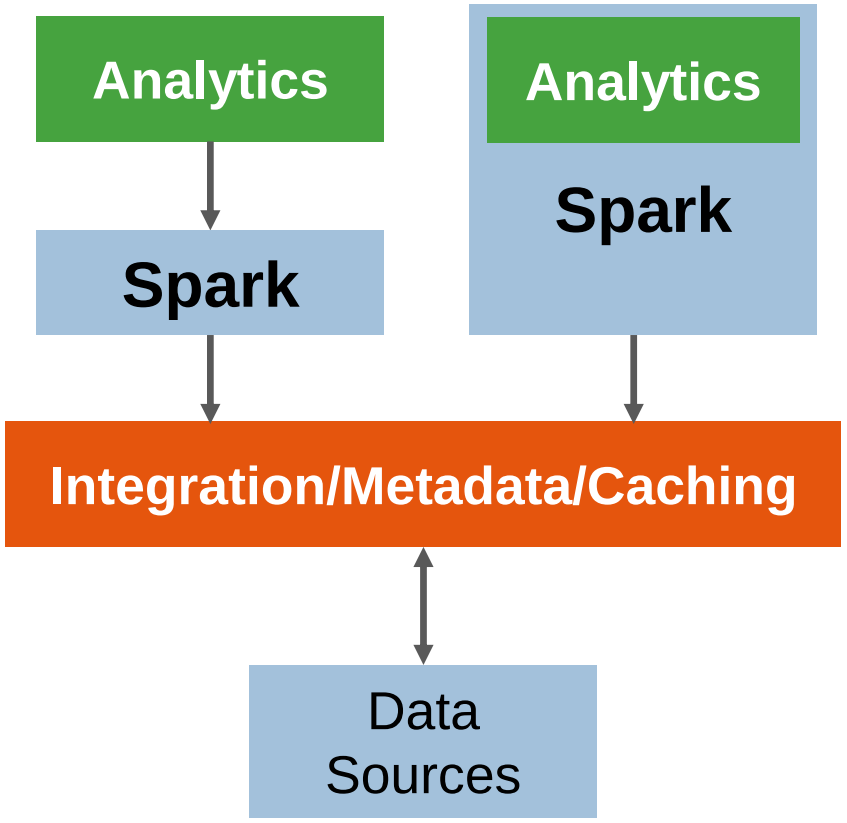
Single-Tier Architecture/ Local Storage



Two-Tier Architecture/ Distributed Storage



Three-Tier Architecture/ Distributed Storage

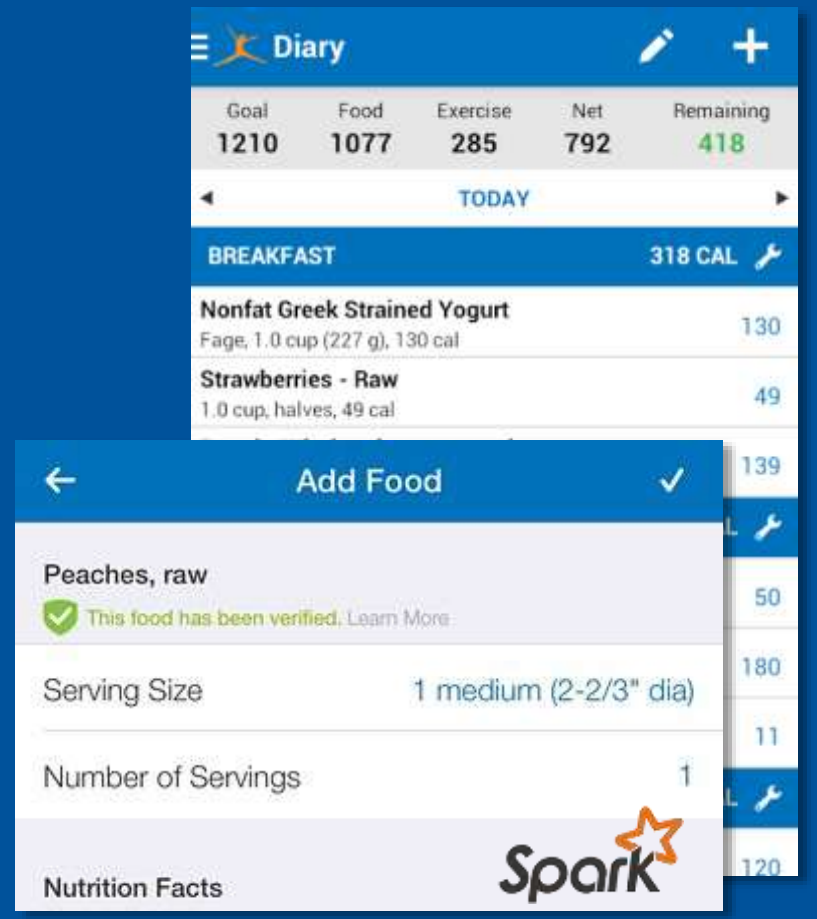


Logical Data Warehouse Architecture

Gartner

Data Preparation and Integration at MyFitnessPal

- The Challenge:
  - 80 million users add >20 million food entries each day, but user-contributed data is messy and inaccurate
- Solution:
  - Hadoop and MapReduce took days to process:
    - Hadoop still used for bulk ETL tasks
  - Spark runs in AWS, using S3 and Databricks
- Benefit:
  - Spark's memory-centric characteristics allow 12-person team to automate validation and correction of 2.5TB of data in minutes



Gartner

Key Issues

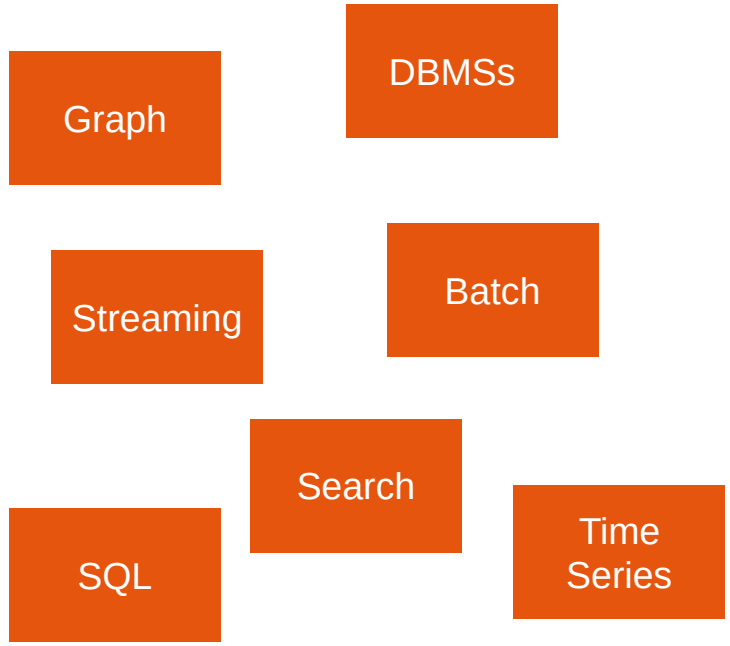
1. What are the use cases for Hadoop and Spark?
2. How will these technologies evolve over the next three to five years?
3. How do you prepare for an uncertain future for Hadoop, Spark and emerging technologies?

Evolutionary Trends

- Emergent:
  - Hadoop, Spark, Kafka usage disaggregate into constituent project adoption
  - Cloud redraws competitive map, highlights separation of compute and storage
- Sustaining:
  - More vendors commercialize open source "Hadoop- or Spark-based" projects not supported by today's distributors
  - Security as a product differentiator heats up
  - Pace of change and proliferation confuse and stymie users

Emergent Trends: Disaggregated Use of Components

- Buyers want parts of stack, add other pieces:
  - Composable stacks become the norm
- Data in motion separates from data at rest:
  - Hadoop at the edge, the center, and in between
  - Cloud drives increased end-user demand
  - Vendors monetize additional offerings: e.g., Hortonworks moves Kafka to HDF



Hype Cycle for Data Management, 2017 calls Hadoop Distributions "obsolete before plateau." This is why.

Gartner

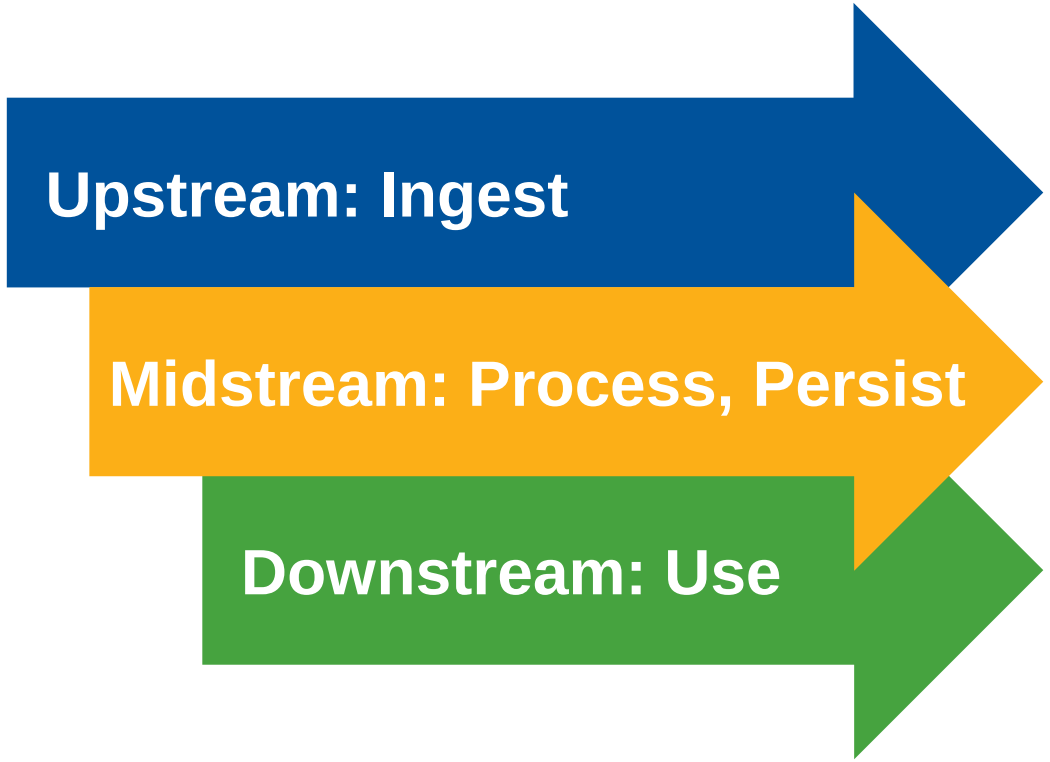
Emergent Trends: Cloud Redraws the Map

- Hadoop in the cloud:
  - Amazon EMR — the first, and largest, steps up its pace, adds pieces
  - Microsoft — only offers Hadoop in cloud today, connects on-premises
  - Oracle Big Data offerings — cloud (and on-prem. partners, appliances)
  - IBM abandons Open Platform, resells/supplements Hortonworks
  - Google Cloud Platform and BigQuery draw attention
- Spark in the cloud? Databricks, Azure, Google …

Not all workload prices benefit from cloud models — evaluate continuous vs. intermittent and model pricing accordingly

Sustaining Trends: Security, Governance as Differentiation, but Securing Big Data Pipelines Spans Multiple Stages

- Distributors build and partner for their own stacks:
  - Cloudera: Sentry, etc.
  - Hortonworks: Atlas, etc.
  - MapR: Access controls
- But their coverage is limited
- Third-parties fill the gaps:
  - Incumbents: Informatica, others
  - Specialists: Blue Talon, others

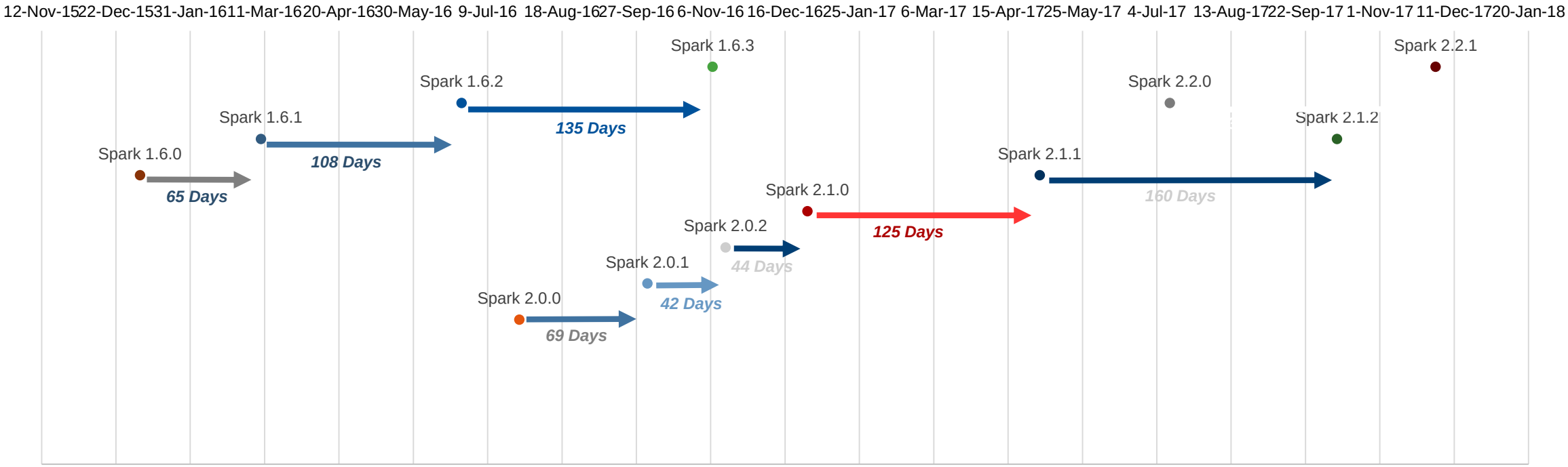


Data Security Threats and Responses Vary

	Ingest	Process	Use
Issues in Big Data Analytics	▪ Data ingested multiple times, but with different permission ▪ Multiple records with slightly different data from the same person, event, etc. ▪ Schema on read	▪ Direct attacks, adversary can manipulate ("poison") training data, parameters, learning algorithm ▪ Infrastructure attacks	▪ Adversary can infer missing or restricted data ▪ Indirect attacks, an adversary can use intersection of intermediary results or out of channel attacks
Controls and Ramifications	▪ Detect replicated data (data discovery) ▪ Data classification and tagging (metadata) ▪ SDM ▪ Detect data relationships at ingest ▪ Programmable ELT ▪ Data watermarking	▪ Limit adversary knowledge ▪ SOD: ETL, workflow and learning, analysis (Atlas) ▪ Stringent permission management ▪ Encryption ▪ Workflow management	▪ Preprocess queries ▪ Triggers and DAP to ensure that queries conform with security policy ▪ Query history retention and analysis ▪ DDM ▪ AuthNZ/Kerberizing ▪ Encryption ▪ Audit + record service

Sustaining Trends: Pace of Change and Proliferation Confuse and Stymie Users

2 years of Spark releases



Releases arrive too quickly to absorb, update and operate

New options: data lakes, event processing, accelerators, operations, HaaS

Gartner

Key Issues

1. What are the use cases for Hadoop and Spark?
2. How will these technologies evolve over the next three to five years?
3. How do you prepare for an uncertain future for Hadoop, Spark and emerging technologies?

Gartner

Elements of Uncertainty

- Evolving vendor landscape:
  - Megavendors continue embrace and extend strategy, pressuring pure plays
  - Cloud plays offer agility and cost optimization over static on-premises deployments
  - Open-source business models haven't been proven
- Dynamic software environment:
  - New platforms and frameworks continue to fragment developer market and mind share
- Improvements in hardware:
  - Optical to the chip, GPUs, increasing memory density and reliability

Evolving Apache-Based Vendor Landscape Adds to Confusion — "What Is Hadoop"?

- data Artisans Flink — "stateful stream processing"
- DataTorrent Apache Apex — also streaming
- Apache Beam — streaming SQL API unification
- Dremio — Apache Arrow-based data lake enablement
- Kyligence — enterprise OLAP on Hadoop
- Hortonworks Apache Metron for cybersecurity
- Arcadia — Data Apache Spot cybersecurity visualization

and many others …

Gartner

Making Your Landscape Less Uncertain — Focus Areas

Component — not core

Operational and analytics skills

Security and metadata

Gartner

Recommendations

- ✓ Use Hadoop and Spark for new workloads leveraging their respective strengths — they are still not a "replacement" story.
- ✓ Define data governance strategies before you begin.
- ✓ Plan for a rapid pace of change in framework components and weigh that against your expectations.
- ✓ Create and stay with a planned release schedule — just as you do with other data management software. Do **not** let vendors dictate it.
- ✓ Cloud isn't right for everything — evaluate costs for continuous vs. intermittent workloads.

Gartner

Recommended Gartner Research

- **Toolkit: Answers to the FAQs on Hadoop Infrastructure**
  Arun Chandrasekaran, Merv Adrian and Nick Heudecker (G00311202)
- **Rethink and Extend Data Security Policies to Include Hadoop**
  Merv Adrian (G00298911)
- **Market Guide for Hadoop Operations Providers**
  Merv Adrian and Others (G00301458)
- **What Apache Spark Means for Big Data**
  Nick Heudecker (G00271327)
- **An Introduction to and Evaluation of Apache Spark for Big Data Architectures**
  Sanjeev Mohan (G00324340)

For information, please contact your Gartner representative.

Gartner