CLOUD COMPUTING APPLICATIONS

BIG DATA PIPELINES: THE MOVE TO HADOOP

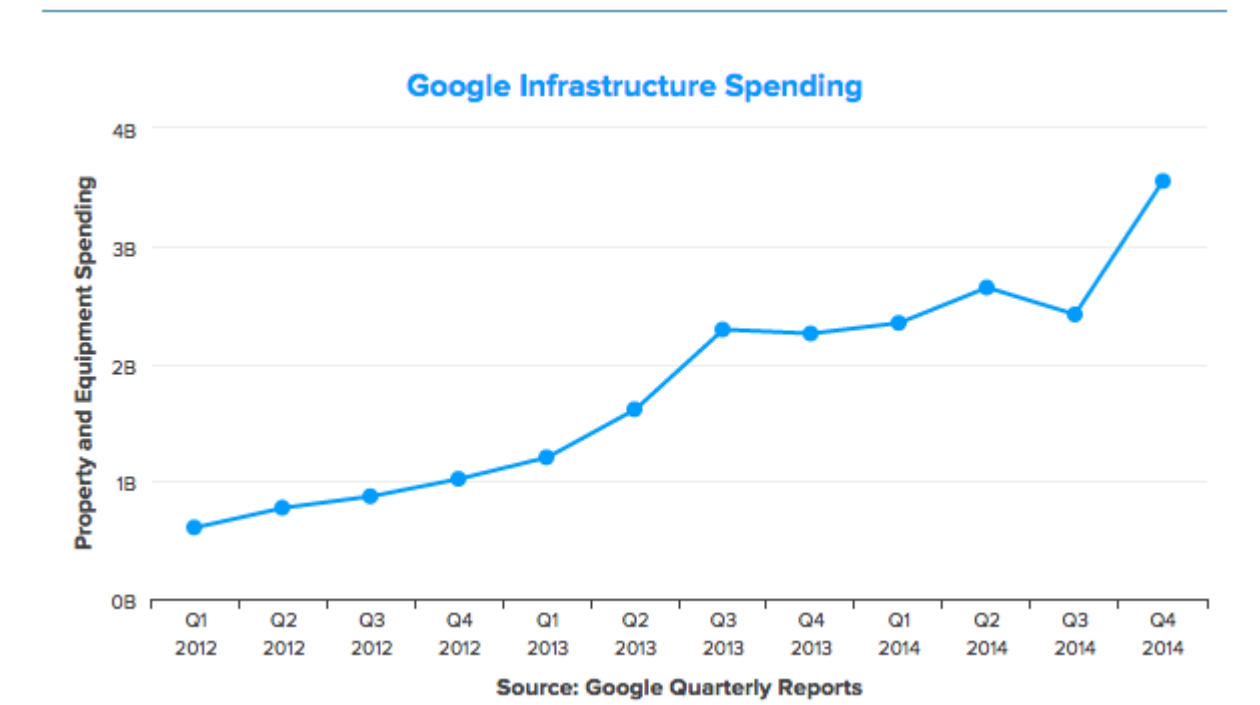
Matt Ahrens – Yahoo

Why Pipelines Are Behind Everything

- With the rise of large data sets, there needs to be a system that can reliably and quickly organize the data
- “Big data” is the trend in the industry, but how do you actually obtain data that is useful on a regular basis?
- Use cases
  - Relevant content tailored to users
  - Programmatic digital advertising
  - Data analytics for research and sciences

Why Pipelines Are Behind Everything

- Data keeps growing



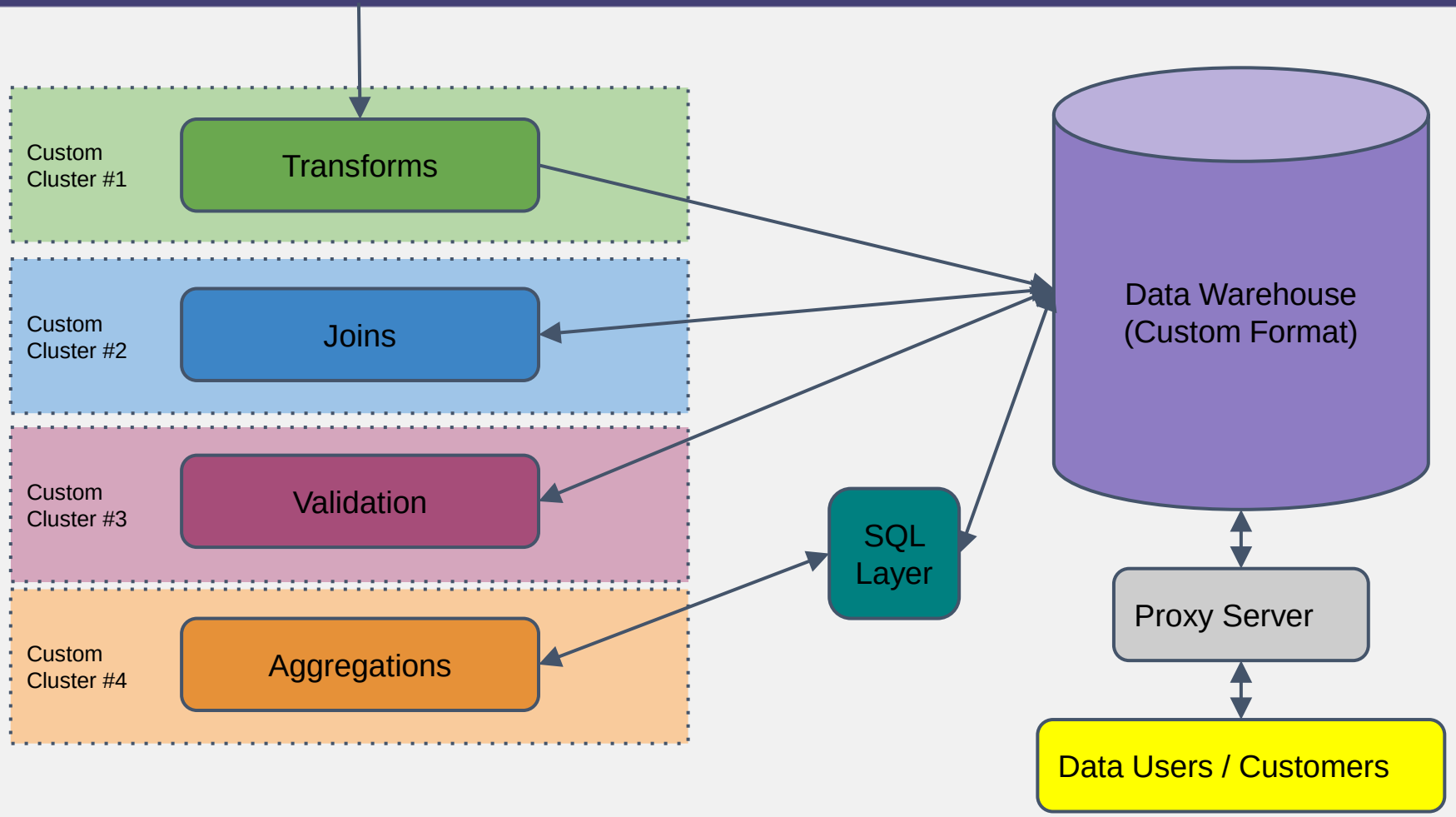
What Is a Data Pipeline

- Simple definition: system that transforms events into a usable format
- Input: raw logs, interactions, activities
- Output: data sets for specific users (filtered, aggregated, joined, etc.)
- Data size scale
  - Billions of transactions per day (millions / minute)
  - TBs of data per day (GBs / minute)

Where We Came From

- Customized mini-clusters of hardware
  - Tailored to specific type of job: transformation, joins, aggregation
  - Pro: mix of memory/cpu config specific for job type
  - Con: scale issues, overhead of HW setup/maintenance
- Lack of well-defined interfaces and APIs
  - No standard schema format or data model
- Data access limitations
  - Access was limited to core developers with advanced data and programming knowledge

The Past Architecture

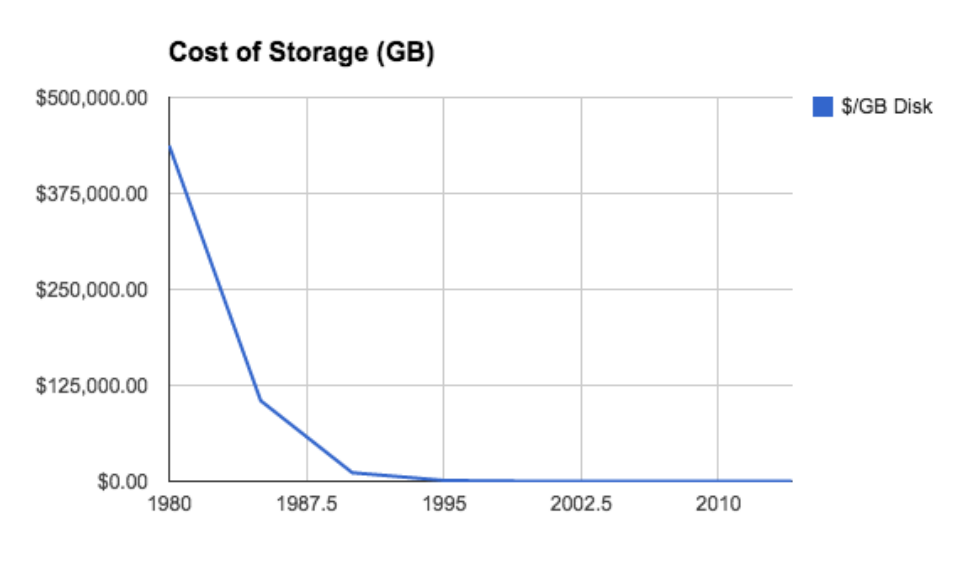


The Elephant Comes Into The Room

Why Move To Hadoop?

- Legacy systems were not performing well (< 1 TB / day)
- We had customers who wanted access to raw feeds (TB / day per customer)
- The advertising roadmap called for a 3-5x increase in traffic (new features, new customers onboarding)

Year	$/GB Disk
1980	$437,500.00
1985	$105,000.00
1990	$11,200.00
1995	$1,120.00
2000	$11.00
2005	$1.24
2010	$0.09
2013	$0.05
2014	$0.03

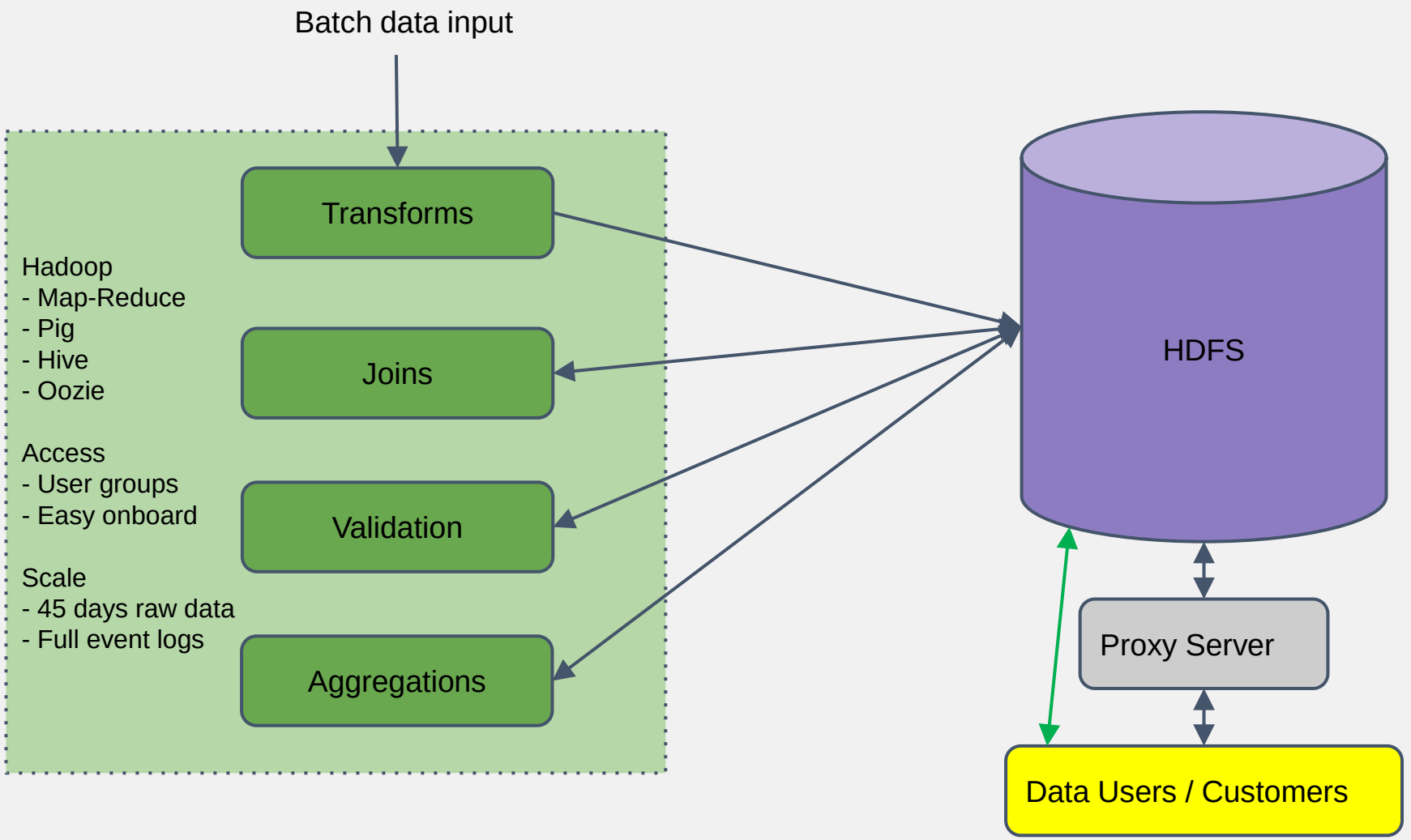


The Promise of Hadoop

- PB+ storage capabilities
  - Multi-tenant internal clusters made up of 1000s of nodes could handle TB/day easily
  - Storage was fault-tolerant with default 3x replication
  - Easy to scale up as new growth occurred

- Hosted service for job execution and data storage
  - No more need for separate clusters as Map/Reduce could handle all types of jobs
  - ETL operations easily handled using Pig Latin interface
  - New innovative frameworks were starting up (HBase, Hive, Oozie) promising more platform adoption

The Architecture on Hadoop



Life on Hadoop

- Platform hardening had its consequences
  - Migrating data users and customers to the new system took longer than expected
  - Running large-scale data pipelines on multi-tenant clusters caused customer issues
- Data for everyone (who is permitted)
  - Number of data users increased dramatically on Hadoop
- Scaled better than expected (over past 5 years)
  - As data size has continued to grow, job runtime and data latency has continued to shrink