Infinite Galaxy!

Brian Finley, Principal Architect, Big Data

2016.06.28
Agenda

1. Who’s that Bald Guy?
2. The Proposition
3. Spark?
4. Architecture
1992
The Proposition

- Incorporate Spark into the Galaxy ecosystem
  - Support Spark enabled tools
  - One tool at a time

- Spark enabled tools scale from your notebook to a cluster with no further tool code changes
Spark – what is it?

- Open Source Framework / Engine
  - Fast with interim results in memory on multiple nodes
  - Easy to write Spark enabled code
  - Parallelism without being a parallel code expert

- Native libraries for many languages:
  - Python
  - Java
  - Scala

Lightning-fast cluster computing
Spark and Big Data

- Outgrowth of the Big Data paradigm
- Hadoop at the center of the Big Data universe
- Designed to work with or without Hadoop
The Architecture Whiteboard

Galaxy Node

Galaxy

Filesystem
The Architecture Whiteboard

Galaxy Node

Galaxy

Spark

Filesystem
The Architecture Whiteboard

Galaxy Node

Galaxy

Spark

Filesystem

As simple as NFS
The Architecture Whiteboard

Galaxy Node

Galaxy

Spark

Spark Node

Spark

Filesystem

As simple as NFS
Spark with Hadoop

- HDFS is great, but…
  - API for filesystem access
  - Greater initial effort for Galaxy

![Diagram of racks](image-url)
thanks.