Big Data Use Case

How Rackspace is using Private Cloud for Big Data

Bryan Thompson

May 8th, 2013

Our Big Data Problem

- **Consolidate all monitoring data for reporting and analytical purposes.**
  - Every device (server, switch, SAN, UPS, etc.) and product produces multiple events per second
  - Monitoring tens of thousands of devices (both physical and virtual)
  - This adds up to terabytes of data per day, and growing…

Current Environment

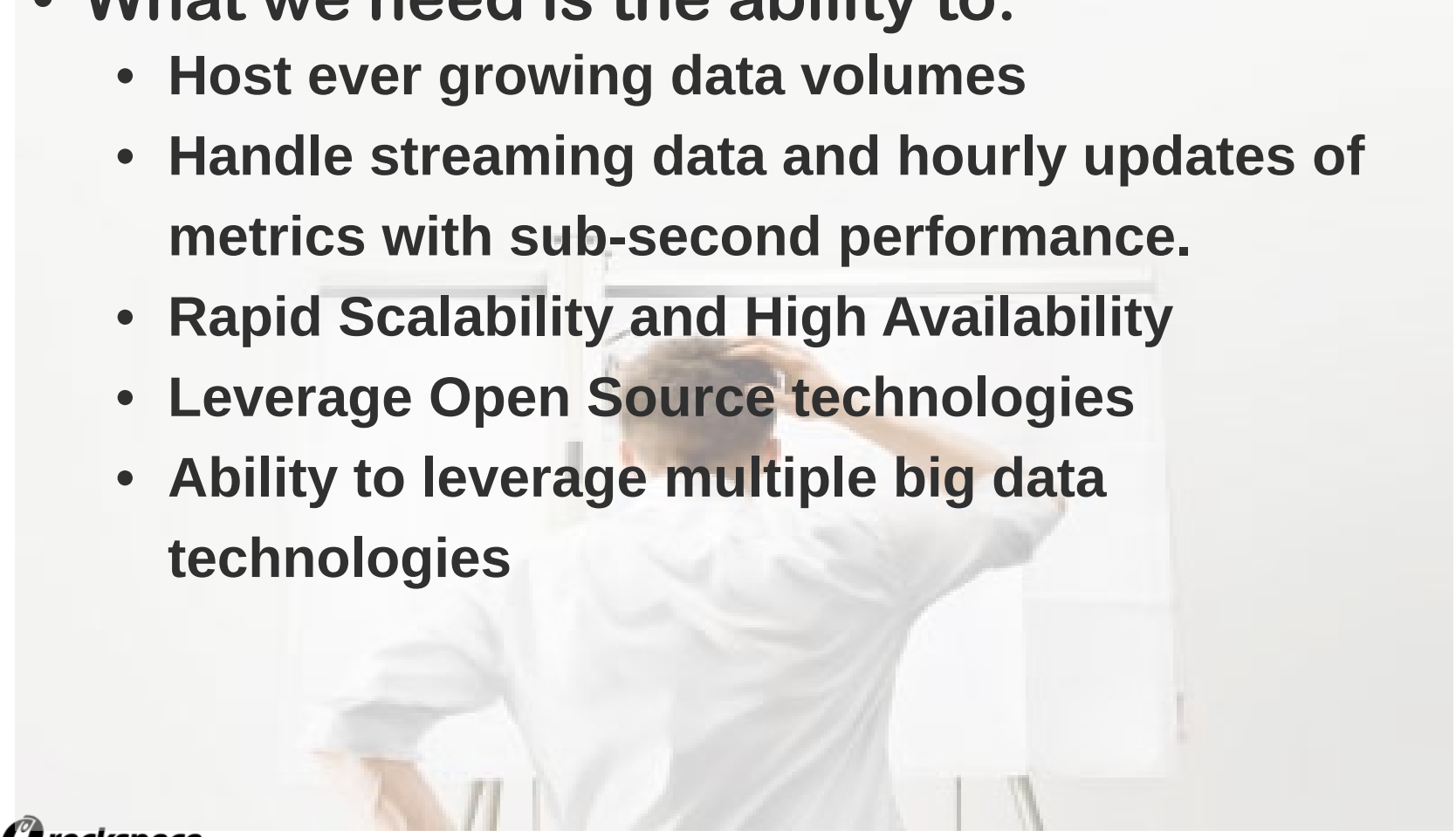
- **Dedicated Relational Database systems**
- **Loaded nightly**
- **Multiple BI Tools**
- **2450+ Users**
- **To scale would be cost and time prohibitive:**
  - Cost of DB licenses
  - Cost of Hardware
  - Time to procure and configure servers
  - Concerns with performance
  - Heavy DBA work

What our sponsors and end-users want…

- Plug in and start analyzing data
- Act at the speed of the business
- Maintain optimal query performance
- ↓ Costs to store and analyze ↗ Data Volumes
- Abstract technical nuances of multiple big data technologies
- Use your preferred BI tool
- High Availability

To The Drawing Board!!!

- What we need is the ability to:
  - Host ever growing data volumes
  - Handle streaming data and hourly updates of metrics with sub-second performance.
  - Rapid Scalability and High Availability
  - Leverage Open Source technologies
  - Ability to leverage multiple big data technologies

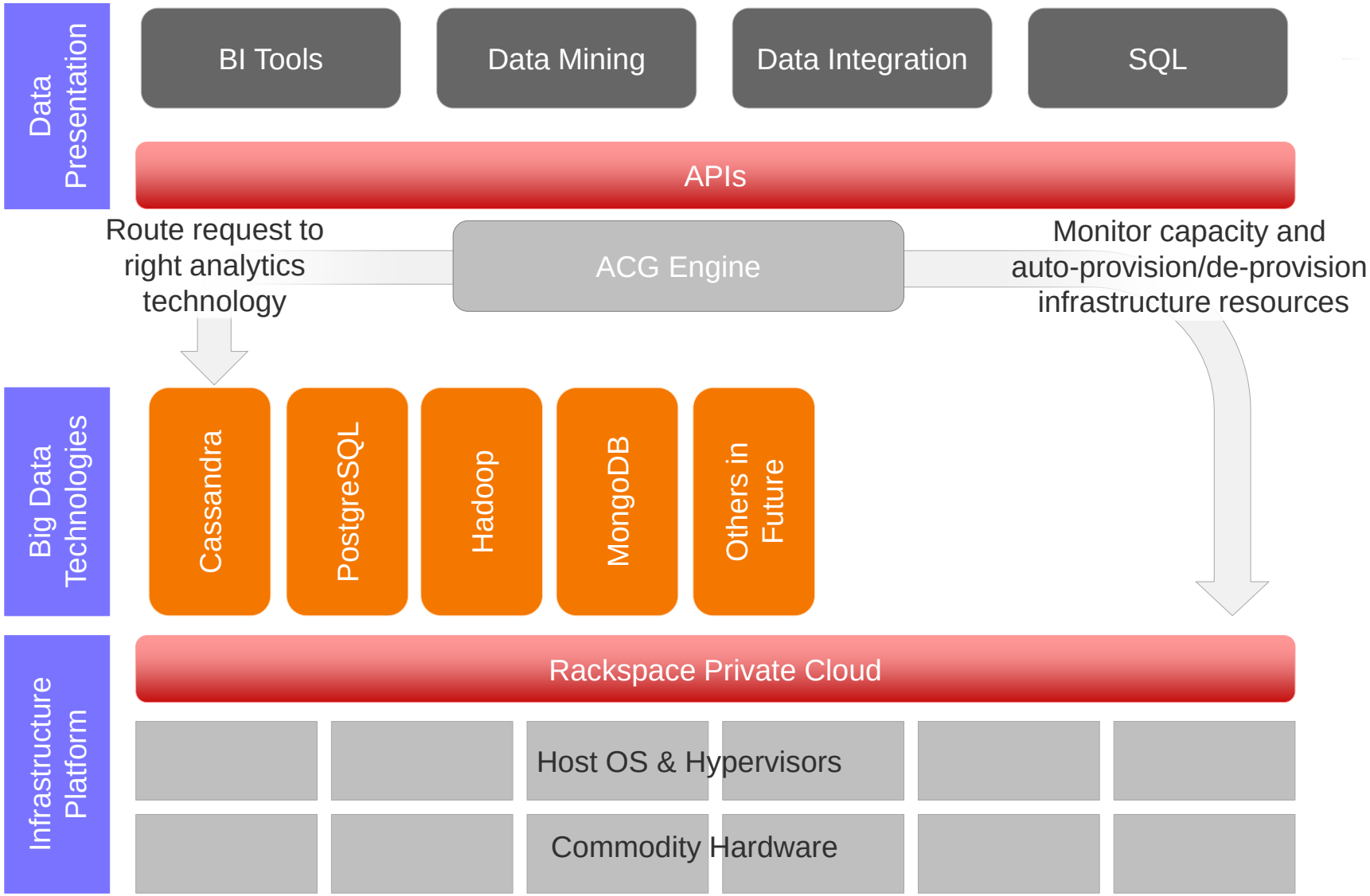


The Analytic Compute Grid (ACG)

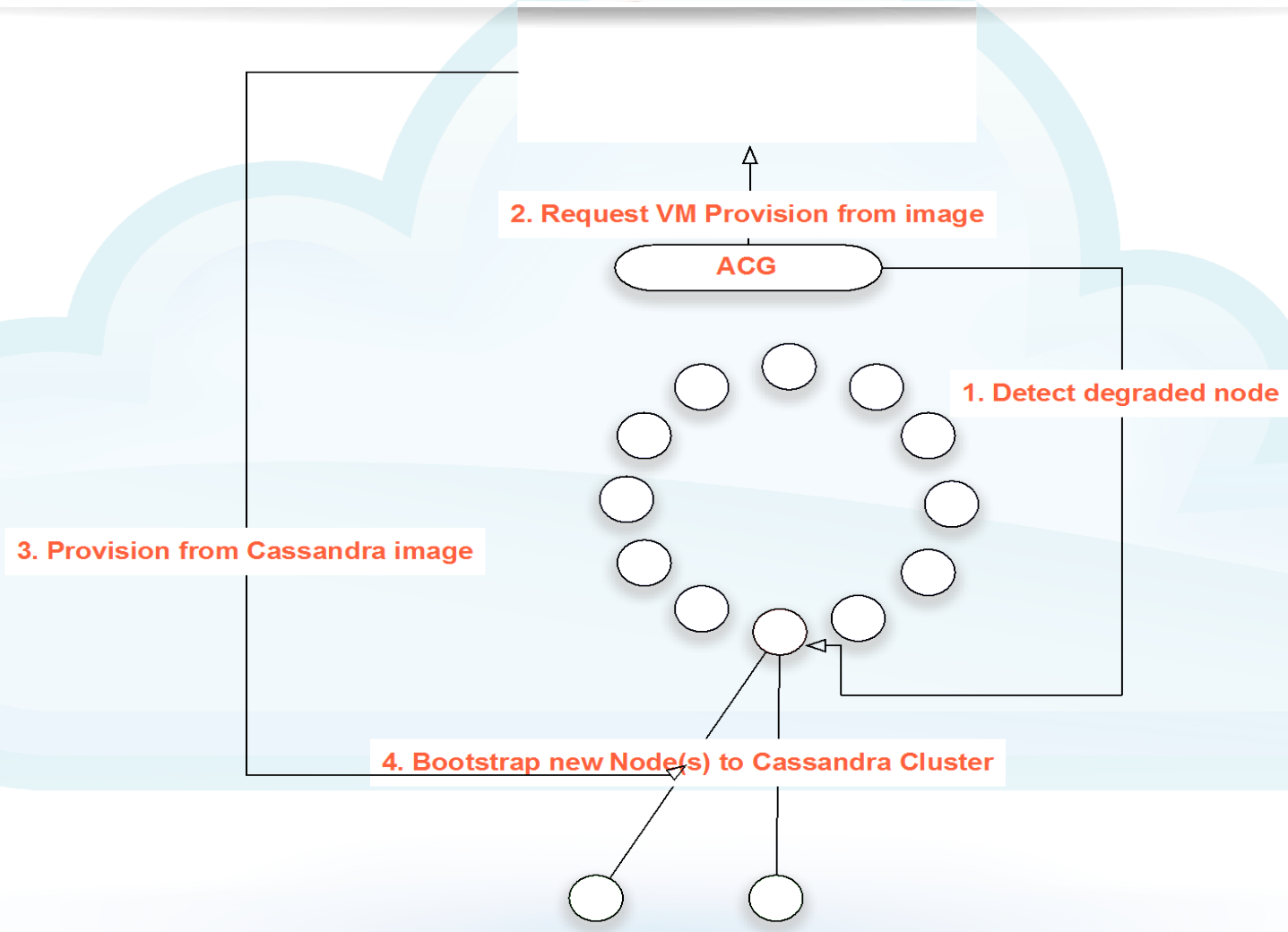
Key components of the ACG

- ✓ OpenStack can provide elasticity capabilities
- ✓ Big Data Technologies (v1 Cassandra)
- ✓ Advanced Hashing to run parallel clusters
- ✓ Rule-based elasticity engine integrated w/ OpenStack
- ✓ ANSI-SQL API w/ Extensions – ability to “plug in”

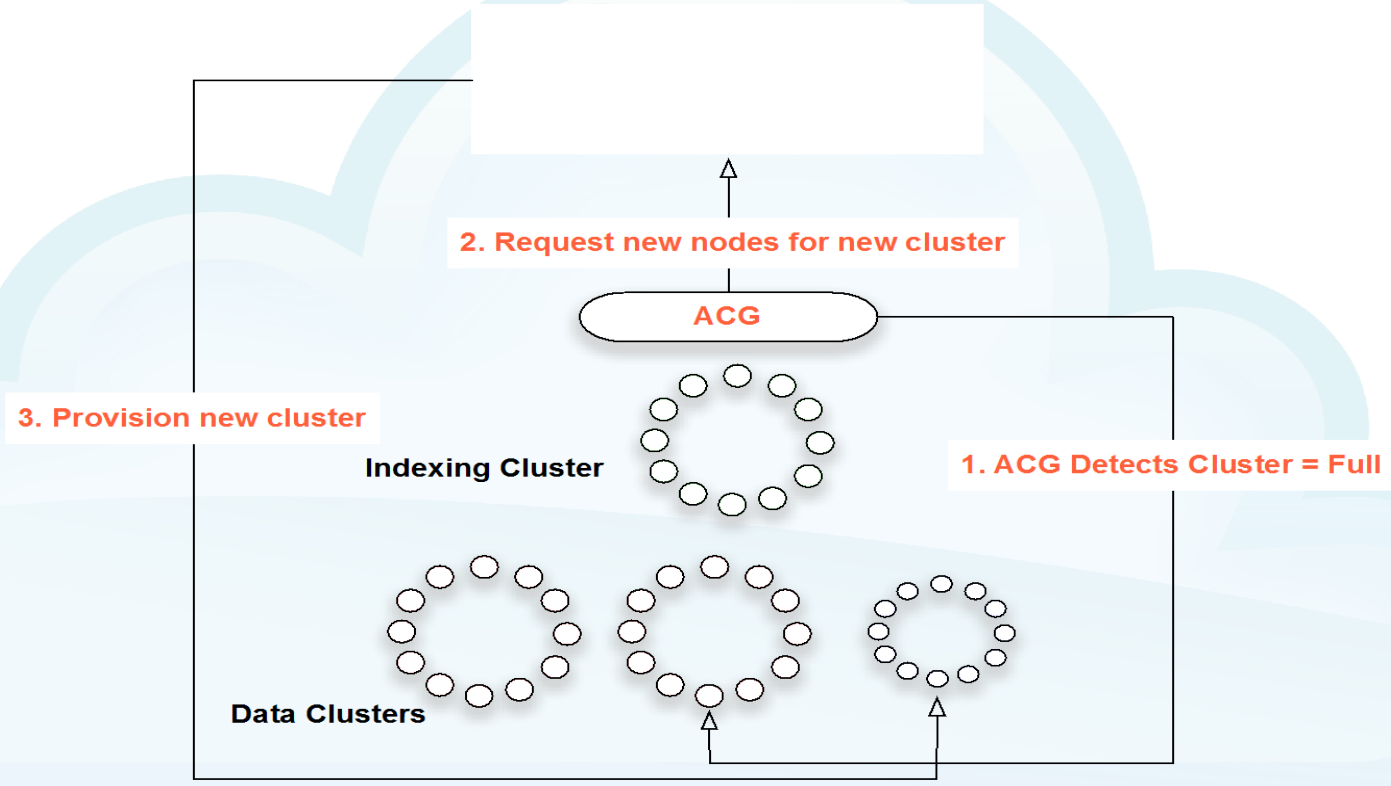
ACG Architecture



Rule-Based Elasticity



Rule-Based Elasticity



Our OpenStack Environment at Launch

- Deployed on Rackspace Private Cloud
- Can run multiple node configurations
- New node is provisioned in seconds!!!
- Operating System – Ubuntu
- Big Data Technology – Cassandra
- 32 Node Cluster – with capacity to grow

Performance Comparison

- **SQL Server Environment (Dedicated Environment)**

24 CPU

256 GB RAM

Availability Calculation against 1.5 Billion row sample – 132 hours (5.5 days)

Performance Comparison

- **RPC OpenStack Environment – (virtual machines)**

Cassandra

8 CPU

32 GB RAM

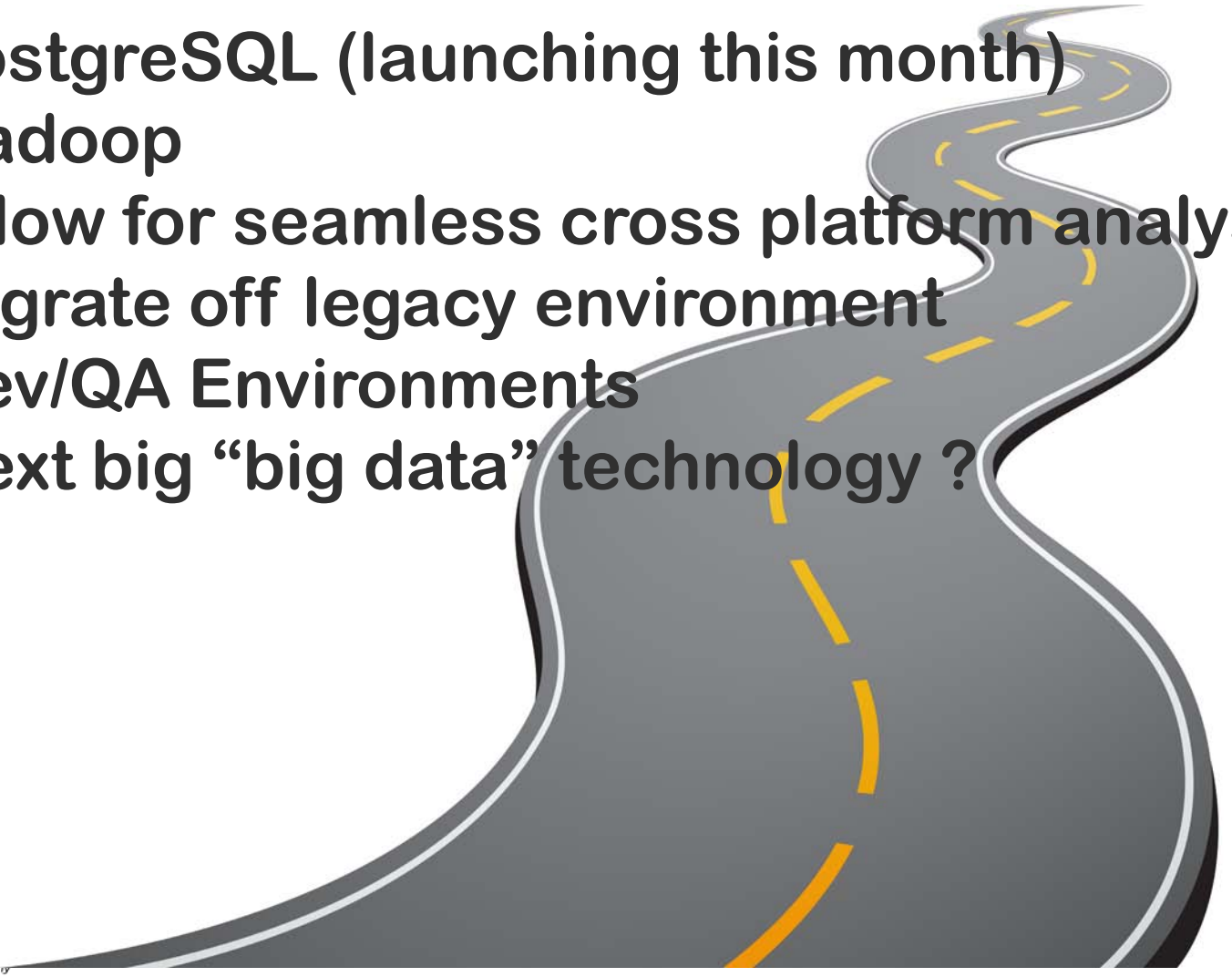
Availability Calculation against 1.5 Billion row sample – 3.2 hours!!!

ACG Features

- **ACG is a Big Data Management System**
- **Parallel engine supports multiple clusters**
- **Highly configurable Rules Engine**
  - Time based
  - System Based
- **ANSI SQL Compliant API with extensions**
- **High Compression - Cassandra**
- **Reusable Bulk-Loader**
- **Can integrate with current ETL tool**

The Road Ahead

- PostgreSQL (launching this month)
- Hadoop
- Allow for seamless cross platform analysis
- Migrate off legacy environment
- Dev/QA Environments
- Next big “big data” technology ?



Questions?

THANK YOU

RACKSPACE® HOSTING | 5000 WALZEM ROAD | SAN ANTONIO, TX 78218
US SALES: 1-800-961-2888 | **US SUPPORT:** 1-800-961-4454 | WWW.**RACKSPACE**.COM