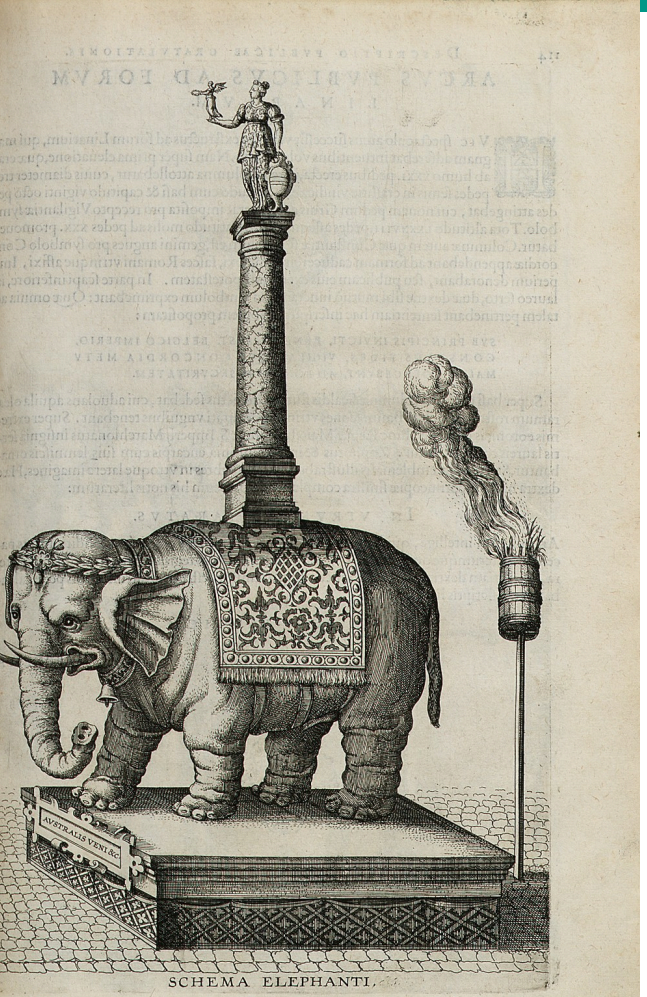


Introduction to Greenplum MPP Database

Andreas Scherbaum

Picture: SCHEMA ELEPHANTI
Author: Jean Boch (Belgian, 1545-1608)
Date: 1595
Book: Descriptio pbvlicae gratvlationis, spectacvlorvm et lvdorvm, in aventv sereniss: Principis Ernesti Archidvcis Avstriae Dvcis Vrgvndiae
Source: Metropolitan Museum of Art
Accession Number: 239 B63 Q
License: Public Domain

Pivotal

Andreas Scherbaum

- Works with databases since ~1997, with PostgreSQL since ~1998
- Founding member of PGEU
- Board of Directors: PGEU, Orga team for pgconf.[eu|de], FOSDEM
- PostgreSQL Regional Contact for Germany
- Ran my own company around PostgreSQL for 7+ years
- Joined EMC in 2011
- then Pivotal, then EMC, then Pivotal
- working on PostgreSQL and Greenplum projects

Agenda

- What is Greenplum
- Architecture Overview
- Live Demo Installation

What is Greenplum

- Massive parallel, shared-nothing Data Warehouse
- Runs on commodity hardware, running Linux
- Analytics Database
- Based on PostgreSQL (was forked)
- Runs basically everywhere:
    - Public/Private Cloud
    - On-Premise
    - Single system or VM
- It’s Open Source!

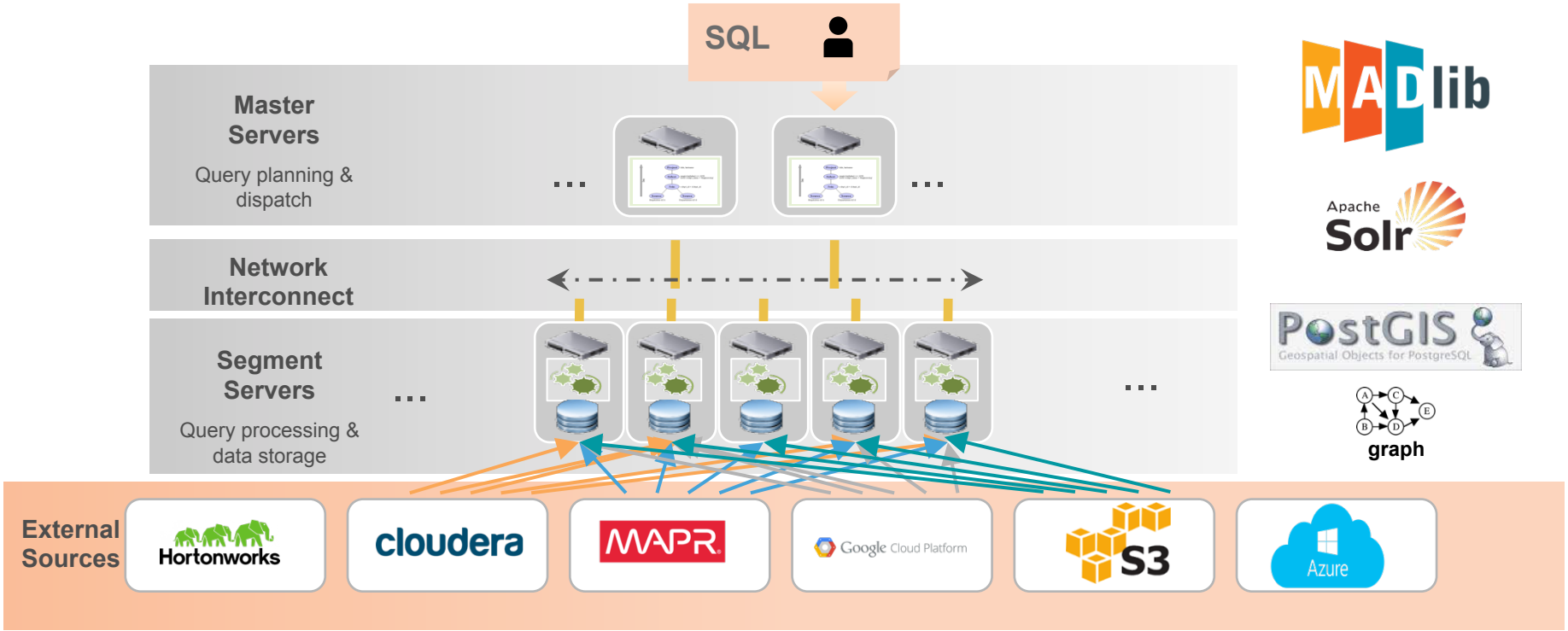
Massive parallel? Shared nothing?

- Massive parallel:
  - Runs on 1-n servers: scales out
  - Several PostgreSQL segment databases per host (segment server)
  - Can scale to several hundred servers, with 1000s of segment databases
  - Biggest known installation: 12 racks a 16 segment servers
- Shared nothing:
  - Servers (and segment databases) do not share anything
  - Separate data storage, network addresses, …

History

- Exists since PostgreSQL 7.4
- Merged with PostgreSQL until 8.2, then forked
- Product evolved, company got acquired
- Open Source in 2015

Architecture



Pivotal

Features

- Workload Manager
- PostGIS on scale
- MADlib for Data Science
- Solr integration
- Command Center for query inspection
- Built-in High Availability and Resilience (Law of Large Numbers)

What problems does it solve?

- Analytics workloads
- OLAP, not so much OLTP
- Large data store (TBs or PBs of data)
- Storage of hot, warm and cold data

Live Demo

- Everything should be automated
    - Even more important if you scale out
- Using Ansible scripts to deploy Greenplum on a single host (VM)
    - https://github.com/andreasscherbaum/gpdb-ansible


THE END