National Aeronautics and
Space Administration

Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

Cloud Computing @ JPL Science Data Systems

Emily Law

Jet Propulsion Laboratory, California Institute of Technology

GSAW 2011

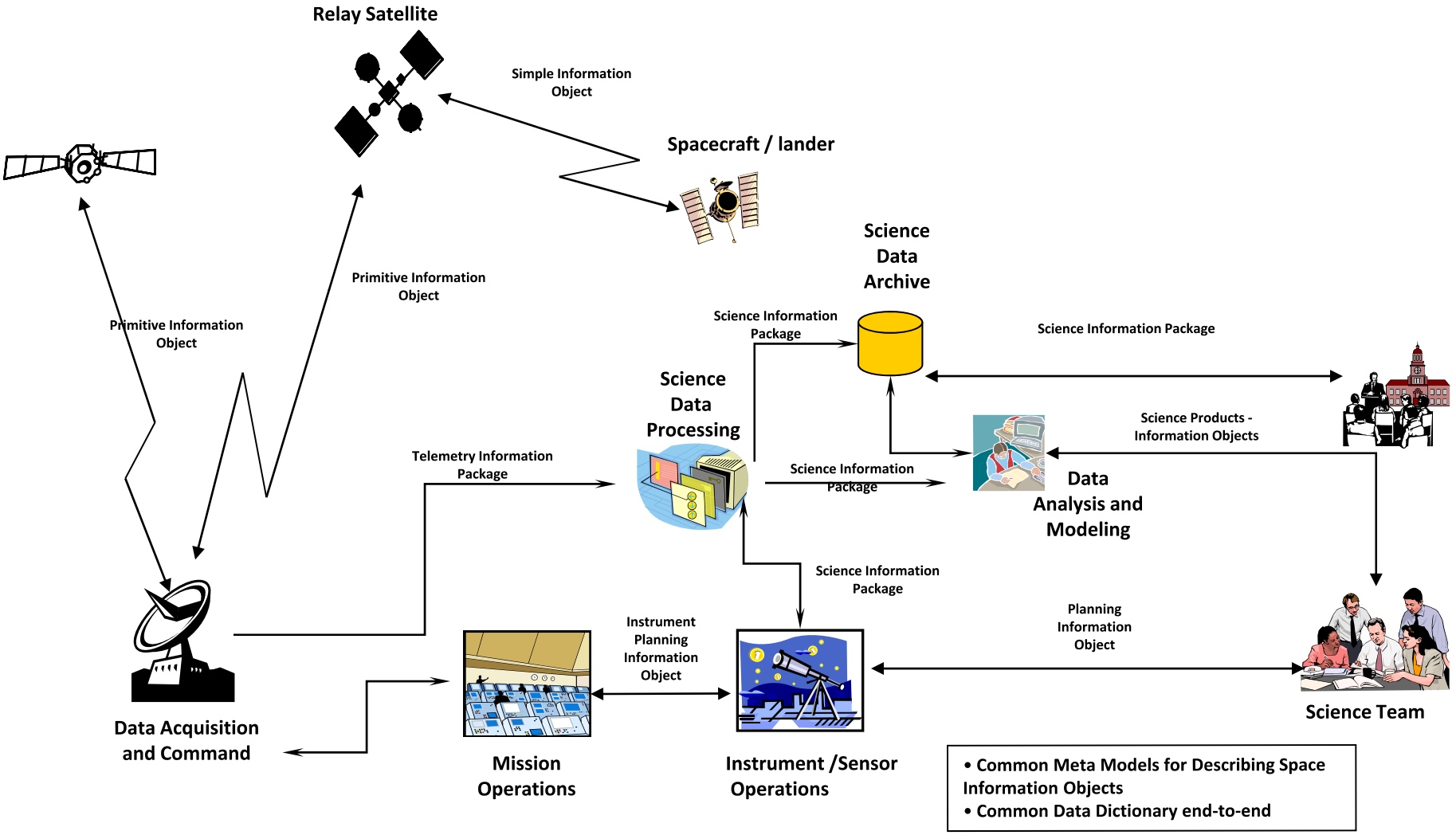
Outline

- Science Data Systems (SDS)
- Space & Earth SDSs
- SDS Common Architecture Components
- Key Components using Cloud Computing
- Use Case 1: LMMP
- Use Case 2: ACCE
- Use Case 3: PO.DAAC
- Conclusion

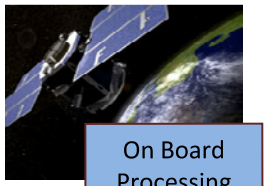
Science Data Systems (SDS)

- Covers a wide variety of domain disciplines
  - Solar system exploration, Astrophysics, Earth science, Biomedicine, etc,…
- Each has its own communities, standards and systems
- But, there is a set of common components
- Some can greatly benefit from proven cloud computing technology

Space Science Data Systems

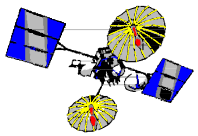


Earth Science Data Systems



SDS Common Architecture Components

Users



Ground Processing Center

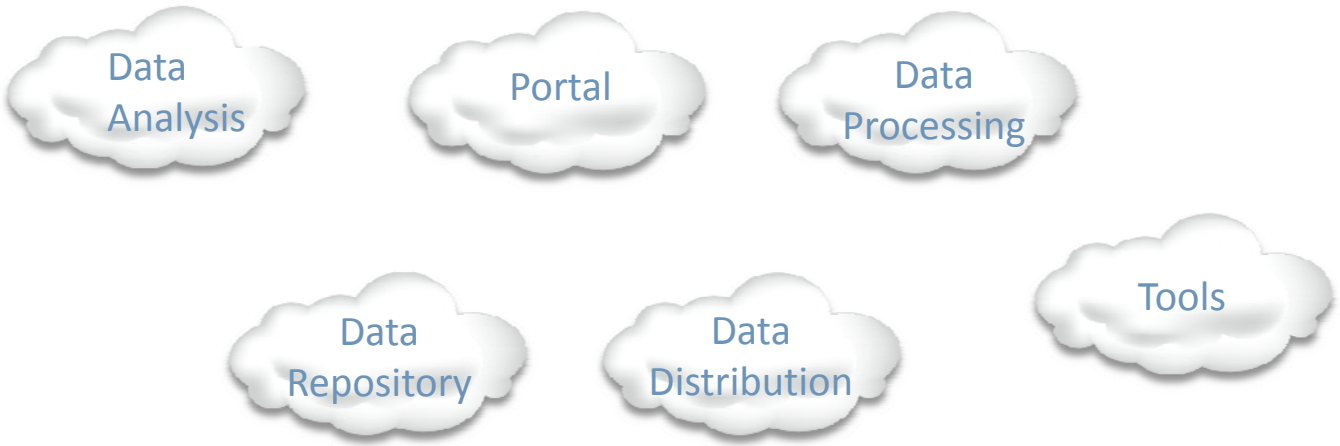
WebService / API

Analysis	Visualization	Tools	Client
Portal			Presentation
Search	Transformation	Transport	Application Service
Data Processing			Data Processing
Ingest	File Manager	Resource Manager	Data Distribution	Data Management
System Resources	Catalog Metadata	Data Repository	Resource
Workflow	Tracking	Monitoring	Security	System Service

Legends:

- System Layer
- Data System Component
- External Entity
- Interface

Key components using Cloud Computing

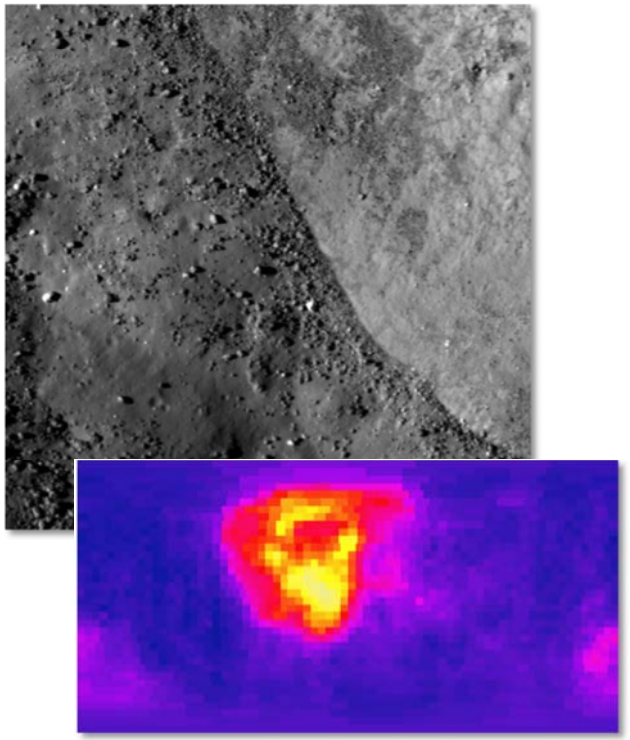


Use Case 1: Lunar Mapping & Modeling Project (LMMP)

- Provides science and exploration community a suite of lunar mapping and modeling tools and products that support the lunar exploration activities
- The tools and products are made available through a common, intuitive NASA portal
- A public release of the system is scheduled at the end of March 2011

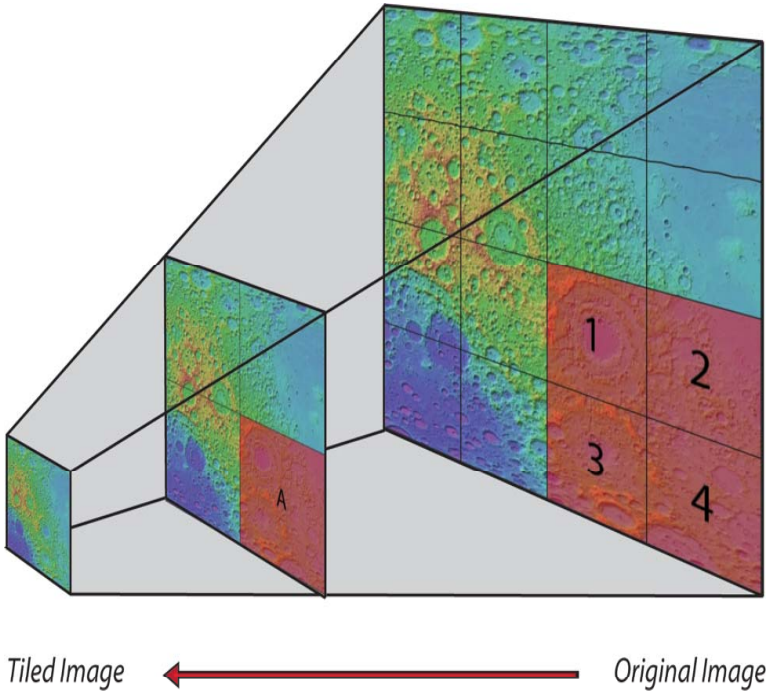
Challenge

- The image files LMMP manages range from a few gigabytes to hundreds of gigabytes in size with new data arriving every day
- Lunar surface images are too large to efficiently load and manipulate in memory
- LMMP must make the data readily available in a timely manner for users to view and analyze
- LMMP needs to accommodate large numbers of users with minimal latency



Cloud Computing Solutions

- Slice a large image into many small images and to merge and resize until the last merge and reduce yields a reasonably sized image that depicts the entire image
- Amazon E2C/S3
- Used distributed approach with Elastic MapReduce to tile images
- Developed a hybrid solution (multi-tiered data access approach) to serve images to users by cloud storage

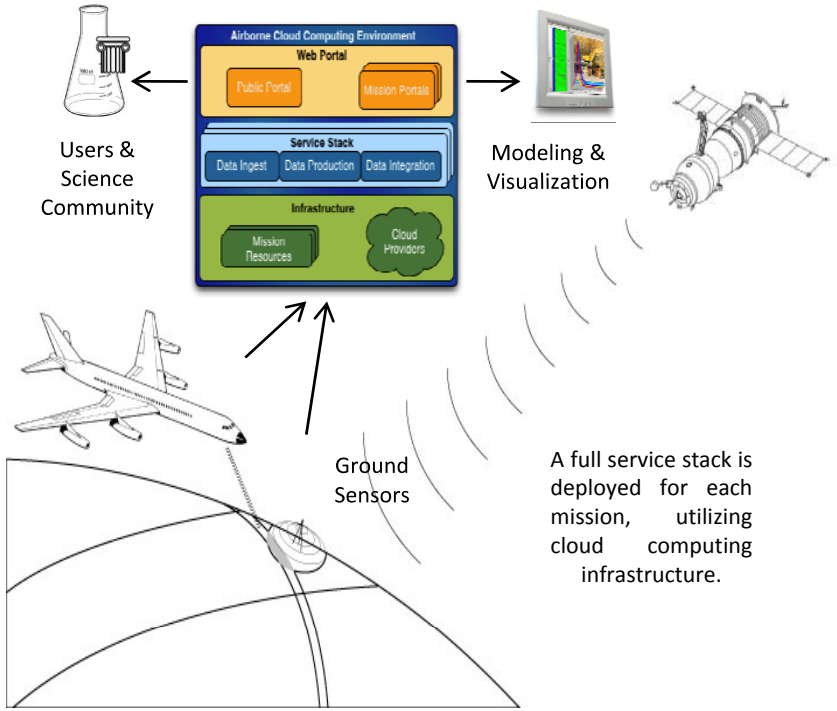


Use Case 1: Results

- Computing performance
  - Comparable especially for the new machines with significant processing capability
  - EC2's "rental" model offers better performance per dollar than having to purchase and maintain local servers
- Storage
  - Pay for just the bandwidth consumed
  - Eliminate the need to purchase extra hardware and bandwidth to handle the occasional spikes in usage
- Security
  - Store only publicly available data on cloud
  - Host private data on local servers to eliminate potential security hole

Use Case 2: Airborne Cloud Computing Environment (ACCE)

- Multi-mission capability providing distributed SDS services applicable to space-borne missions
  - File Management
  - Workflow Management
  - Resource Management
- Extend the existing services to utilize cloud services, commercial, community and private
  - Storage
  - Compute Resources

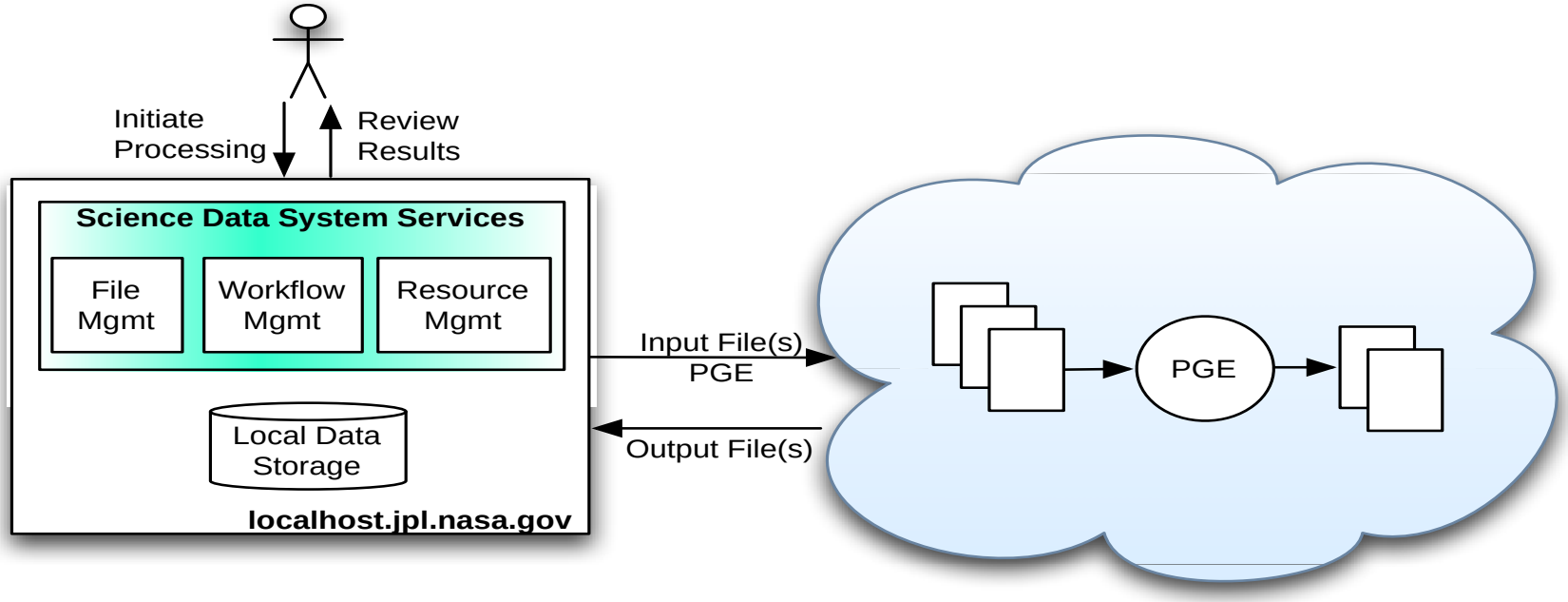


Cloud Computing @ ACCE

- Explore the benefits of performing science data processing for airborne missions in the cloud
- Evaluate different cloud technologies
  - Amazon EC2/S3
    - Elastic compute resources and on-demand storage
  - Eucalyptus
    - Infrastructure software for establishing a private cloud
  - Hadoop – Distributed File System (DFS) and MapReduce
    - Increased processing performance on large data sets

Prototype @ ACCE

A typical science data processing job, Product Generation Executable (PGE), is an executable that performs a task on *n* inputs to produce outputs. The ACCE Cloud must be able find an available resource for the PGE, transfer both the PGE and its input files to that resource, execute the PGE, then retrieve the output files from that resource.



Use Case 2: Results

- Challenges
  - Host Environment
  - Security
  - Network
- Processing cost reduction
  - No investment in capital required (upfront or refresh costs)
  - Pay only for what you use
- Possible limiting factors for public cloud viability
  - Support for ITAR-sensitive data
  - Data transfer rates between JPL and commercial cloud
  - JPL Firewall
- More work ahead
  - Amazon EC2/S3 reported an “ITAR Region” available in Fall 2011
  - Continued benchmarking and optimization has demonstrated increased data transfer rates, 250 – 400 mbps
  - JPL developing a “Virtual Private Cloud” connection to Amazon, causing EC2 nodes to appear inside the JPL Firewall

Use Case 3: Physical Oceanography Distributed Active Archive Center (PO.DAAC)

Preserve NASA's ocean data and make these and related information accessible and meaningful

Cloud Computing @ PO.DAAC

- Understand and articulate the cost/benefit of cloud technologies for the NASA Data Centers
- Evaluate the cloud services provided by Amazon
  - Pricing
  - Data transfer procedures
  - Software deployment and monitoring procedures
- Conduct a pilot study using cloud service for processing oceanographic climatology, data management and data access
  - Utilize MapReduce parallel processing algorithm
  - Benchmark between cloud and standalone processing

Science Approach

- Extract daily time-series and climatologies for different regions over Antarctica (from the cloud)
- Compute climate proxies:
  - Spatial extend, onset and length of melting on the ice sheet
  - Anomalous events
  - Temporal changes/trend in sigma-0
- Examine patterns in the climate proxies and look for correlations with ocean/atmosphere indices

Cloud Computing (Hadoop) solution

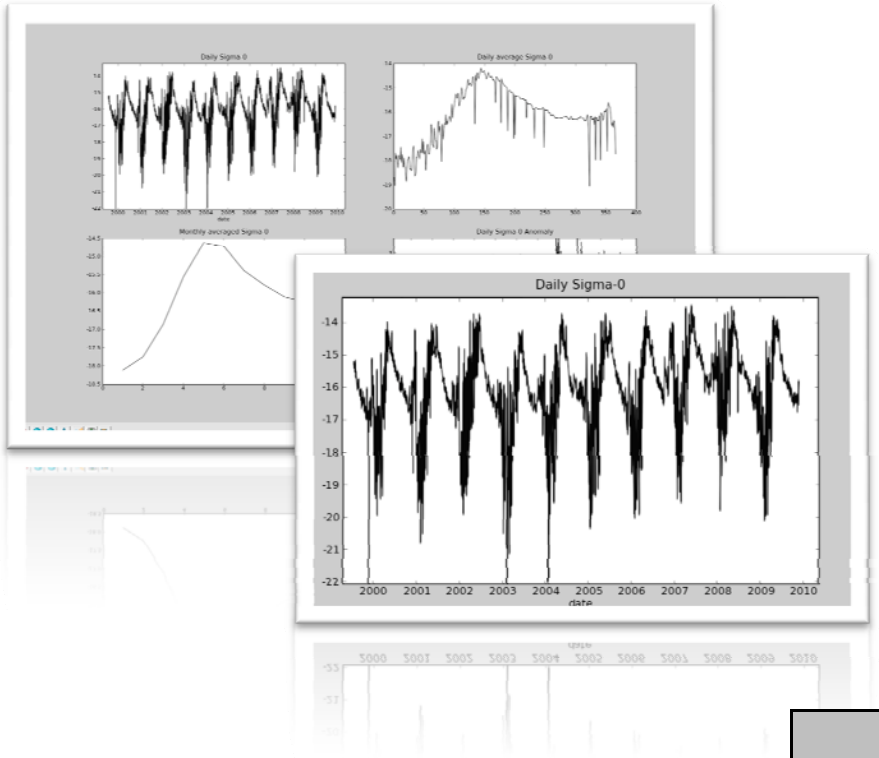
- Copy data into Hadoop HDFS from S3
- 26GB Uncompressed data (550M data points)
- 3x replication

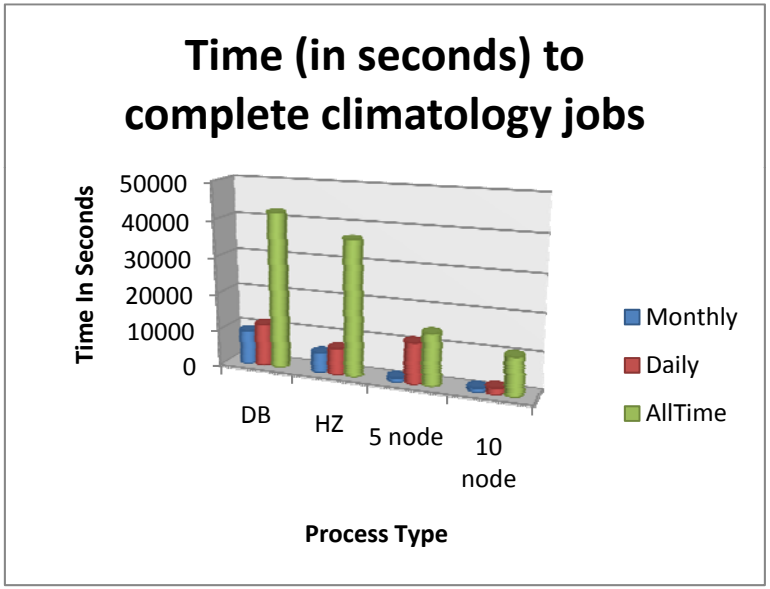
- Compute daily/monthly climatologies
- Store climatology data, all time data points in HBase
- Add point average & standard dev. in future

- Scan ranges of data (lat/long bounding box)
- Aggregate and return average pixel values for given regions

Use Case 3: Results

10 Years at a glance

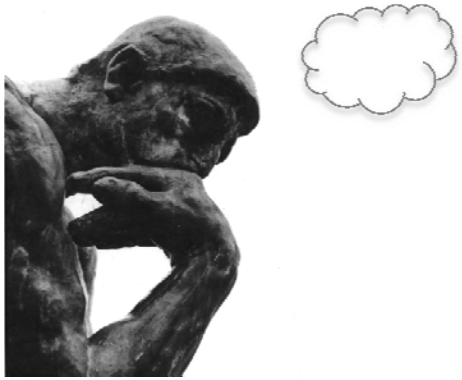




	Oracle	Desktop	5 node	10 node
Monthly	9239	5453	1200	1238
Daily	11502	7169	11204	1848
AllTime	42158	36627	14143	10540

Conclusion

The Silver Bullet?



Many Benefits

Accessible from anywhere

Increase/decrease number of machines based on user defined parameters

Shared pool of configurable computing resources

Resizable compute capacity for unlimited growth
Increase/decrease number of machines based on user defined parameters

Utility Computing, pay by the drink, rapidly provisioned

But….

There are disadvantages to watch out

- Privacy / Ownership
- Security
- Reliability
- Feasibility
- Standards

Thank You

Question