Pipelines

Concepts

Factory Assembly Line

- Manufacturing processes
- Parts from one or more manufacturing processes go into the next manufacturing process
- Some of the manufacturing processes can be done at the same time, that is, in parallel
- End product is a finished good

Pipelines

- Aka data pipelines
- Analogous to a factory assembly line
  - Processes
  - Output from one or more processes goes into one or more processes
  - Some of the processes can be done in parallel
  - End product is data that is ready to go for analytics

Pipeline Goals

- Fully automated—no manual steps
- Handle all data possibilities gracefully
- No crashes—automatically recover from all errors and continue
- Data should be cleansed and validated and ready to go for analytics
- Efficient
  - Data used for analytics often goes stale quickly
  - Need to make data available for analytics as soon as possible
  - Run as much in parallel as we can

Building Pipelines

- Comprehensive list of all input data
  - Encoding, formats, where from, how often, etc.
- Design and build processes
  - Acquire the data
  - Staging load of data
  - Cleanse, validate, transform data
  - Combine with other datasets, especially secondary datasets
  - At the end, data is ready for analytics
    - Not always a load

ETL (ELT): Extract, Transform, Load

- E**T**L (and E**L**T) are subsets of pipelines
- Pipelines do much more
  - Data cleansing
  - Combining with other datasets
  - Parallelization
- Pipelines do not always load data
  - Enterprise message queues, publish, produce, stream, etc.
- More about E**T**L, E**L**T, enterprise message queues in coming weeks

The End

Pipelines

Business Cases

POS (Point of Sales) System

- POS processes transactions
- We cannot wait for the transactions to show up in the data warehouse hours later
- Design a pipeline
  - Stream data to analytics cluster
  - Load data into memory for immediate processing
  - Load data into object store so other systems can have the data for immediate processing
  - Load data into a relational data for later batch processing

Web Server

- Web server generates web logs every time it accesses a web page
- Design a pipeline
  - Stream web logs data to analytics cluster
  - Load data into memory for immediate processing
  - Load data into object store so other systems can have the data for immediate processing
  - Load data into a relational data for later batch processing

Third Party Sales Channel

- Third party sales channel generates sales
- No real-time connection to our sales database
- Presents a daily batch at the end of the day

Third Party Sales Channel (cont.)

- Design a pipeline
  - Wait for the batch
  - Load the batch into staging tables
  - Validate, cleanse, final validation of data
    - Use secondary datasets as needed
  - Combine data with existing data
  - Load data into memory for immediate processing
  - Load data into object store so other systems can have the data for immediate processing
  - Load data into a relational data for later batch processing

The End

Clusters of Containers

Concepts

Vendor-Provided Container Images

- Most software vendors now provide container images for their products
- Easy to create a container from their container image

Single Containers

Suppose

- Several products from several vendors
- Create a single container for all vendor products

Single Container Issues

- Cannot leverage vendor-provided images
- Must load and configure all vendor software in our container
- Cannot scale up by creating multiple containers from the same vendor container image

Cluster of Containers

Solution to single container issues

- Create a cluster of containers
  - Leverage vendor container images
  - Scale up by creating multiple containers from the same vendor container image

Cluster Specification

- Which container images to use for what containers
- How many containers to create from an image
- Startup dependency order
  - Container A needs to be running before container B is started
- Storage mounts
- Networking
  - Hostnames, ports, port mappings, etc.
  - Connections between containers
  - Connections to outside the cluster

One Command Convenience

- One simple command to start up cluster
- One simple command to shut down cluster
- Much easier than manually starting individual containers and networking them together (could take hours)

The End

Clusters of Containers

Business Cases

Database Analytics

- We want to use one vendor for a programming environment and one vendor for the database environment.
- Both vendors have container images.
- Solution: Create a two-container cluster.
  - One container from vendor-supplied database
  - One container from vendor-supplied programming environment
  - Programming container connects to the database container

Streaming Analytics

- We want to use one vendor for our programming environment, one vendor for enterprise message queues, one vendor for relational database, and one vendor for NoSQL database
- All vendors have container images
- Solution: Create a four-container cluster, one for each vendor image, containers connect to each other as needed

The End

Container Orchestration

Concepts

Thinking in Terms of Clusters of Containers

- We start cloud-based VMs (virtual machines) with specifications for CPU, memory, and storage.
- We start clusters of containers in the VMs based on our application needs.
- What if we could specify CPU, memory, and storage at the container level rather than at the VM level?
- Would we care what VMs are needed to make it happen?

Container Orchestration

- We specify clusters of containers
- For each container we specify CPU, memory, storage, etc. typically specified at the VM level
- Container orchestration system will figure out:
  - What VMs are needed
  - What containers to run in what VM
  - How much CPU, memory, storage, etc. for each VM
- We can now just think in terms of containers and do not have to concern ourselves with the VM layer and below

Scale Up and Load Balance

- Start with a minimum number of containers
- Add more containers as demand increase
- Remove containers as demand decreases
- Load balancing: distributing the workload evenly among containers

Cloud-Managed Services

- Cloud-managed services for container orchestration is very popular.
- It frees us from thinking in terms of VMs to thinking in terms of containers.

The End

Container Orchestration

Business Cases

Electric Company Smart Meters

- Electric company with smart meters
- Millions of customers
- During peak usage time, Monday through Friday, 7 AM through 7 PM, we get readings every minute to keep the grid stable
- During off-peak usage times, we get readings every 15 minutes
- All processes are run using containers

Electric Company Smart Meters (cont.)

- Solution: container orchestration
  - Scales up containers during peak times
  - Scales down containers during off-peak times
  - Load balances the incoming streaming data among the containers
  - Need not concern ourselves with the VMs the containers are running in

Similar Scenarios

- POS
- Airline boarding pass readers
- Social media streaming data
- Web server logs

The End