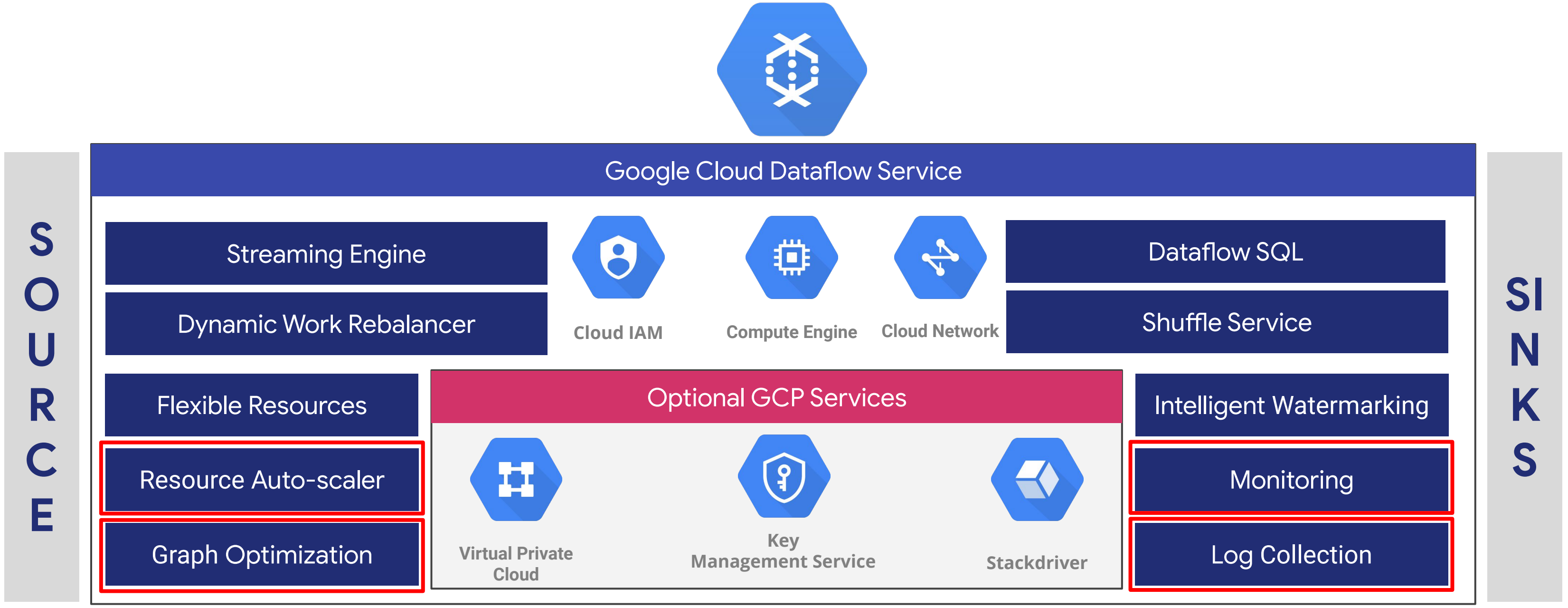
GCP Dataflow Architecture

Svetak Sundhar
Solutions Engineer, Google

Agenda

1. Overview of Dataflow Runner architecture
2. Overview of Dataflow Runner core features
3. GCP Horizontal services integrations
4. New Dataflow Runner features

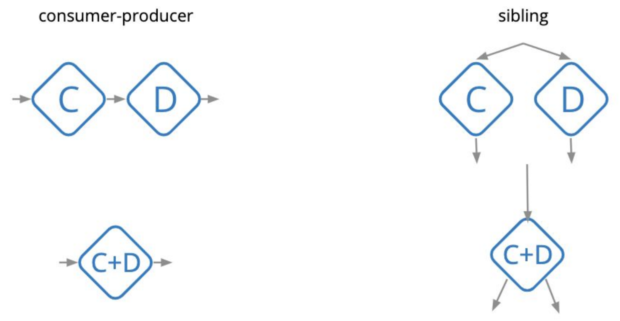
Google Cloud Dataflow Service



Dataflow features

Graph optimization

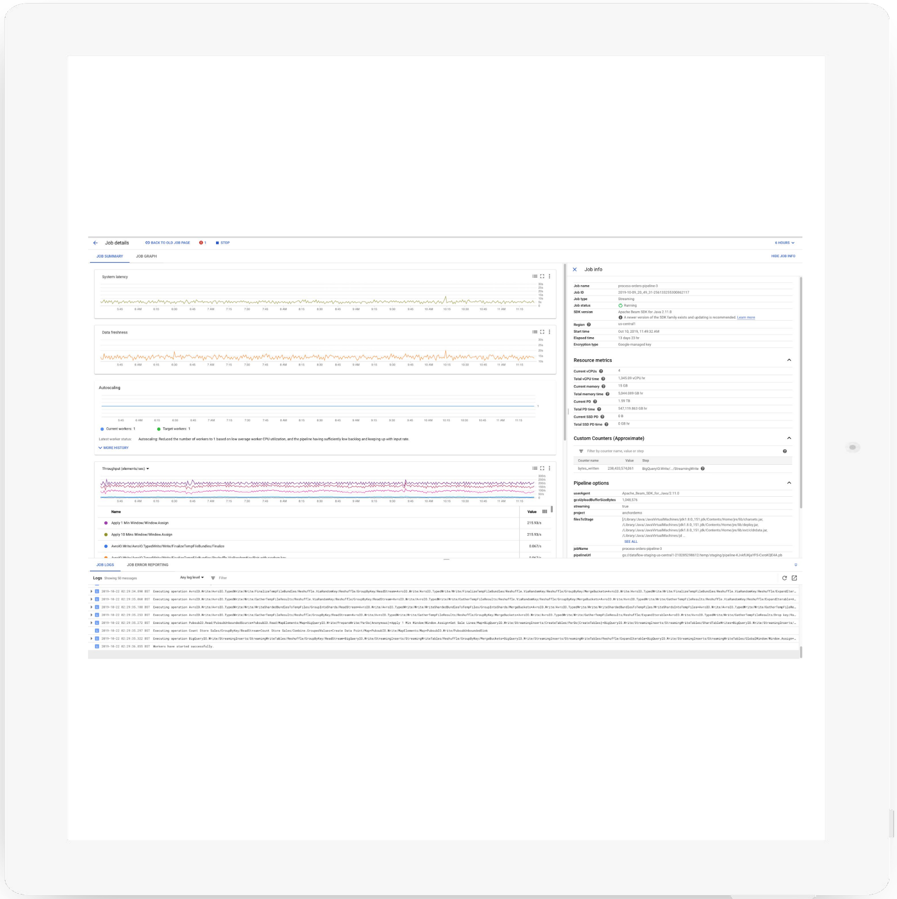
- Producer - Consumer Fusion
- Sibling Fusion
- Others...



Dataflow features

Monitoring

- Dataflow job page
  - Enhanced observability features



Latest worker status: Autoscaling: Reduced the number of workers to 1 based on low average worker CPU utilization, and the pipeline having sufficiently low backlog and keeping up with input rate.

Name	Value
Apply 1 Min Window/Window.Assign	215.93/s
Apply 10 Mins Window/Window.Assign	215.93/s
AvroIO.Write/AvroIO.TypedWrite/Write/FinalizeTempFileBundles/Finalize	0.067/s

Job info

Job name	process-orders-pipeline-3
Job ID	2019-10-09_20_49_31-256133255300862117
Job type	Streaming
Job status	Running
SDK version	Apache Beam SDK for Java 2.11.0 A newer version of the SDK family exists and updating is recommended. Learn more
Region	us-central1
Start time	Oct 10, 2019, 11:49:32 AM
Elapsed time	13 days 23 hr
Encryption type	Google-managed key

Resource metrics

Current vCPUs	4
Total vCPU time	1,345.09 vCPU hr
Current memory	15 GB
Total memory time	5,044.089 GB hr
Current PD	1.59 TB
Total PD time	547,119.863 GB hr
Current SSD PD	0 B
Total SSD PD time	0 GB hr

Custom Counters (Approximate)

Filter by counter name, value or step

Counter name	Value	Step
bytes_written	238,433,574,061	BigQueryIO.Write/.../StreamingWrite

Pipeline options

userAgent	Apache_Beam_SDK_for_Java/2.11.0
gcsUploadBufferSizeBytes	1,048,576
streaming	true
project	anchordemo
filesToStage	[/Library/Java/JavaVirtualMachines/jdk1.8.0_151.jdk/Contents/Home/jre/lib/charsets.jar, /Library/Java/JavaVirtualMachines/jdk1.8.0_151.jdk/Contents/Home/jre/lib/deploy.jar, /Library/Java/JavaVirtualMachines/jdk1.8.0_151.jdk/Contents/Home/jre/lib/ext/cldrdata.jar, /Library/Java/JavaVirtualMachines/jd ... SEE ALL
jobName	process-orders-pipeline-3
pipelineUrl	gs://dataflow-staging-us-central1-210285298612/temp/staging/pipeline-KJvkfUKjaYFS-CxroKQE4A.pb

JOB LOGS　JOB ERROR REPORTING

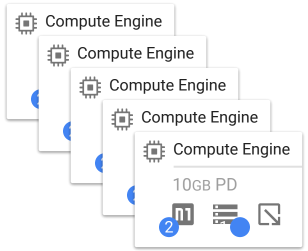
Logs Showing 50 messages　Any log level　Filter

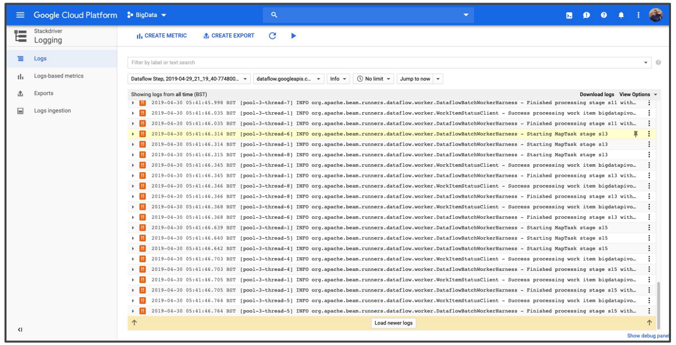
```
2019-10-22 02:29:34.890 BST  Executing operation AvroIO.Write/AvroIO.TypedWrite/Write/FinalizeTempFileBundles/Reshuffle.ViaRandomKey/Reshuffle/GroupByKey/ReadStream+AvroIO.Write/AvroIO.TypedWrite/Write/FinalizeTempFileBundles/Reshuffle.ViaRandomKey/Reshuffle/GroupByKey/MergeBuckets+AvroIO.Write/AvroIO.TypedWrite/Write/FinalizeTempFileBundles/Reshuffle.ViaRandomKey/Reshuffle/ExpandIter…
2019-10-22 02:29:35.060 BST  Executing operation AvroIO.Write/AvroIO.TypedWrite/Write/GatherTempFileResults/Reshuffle.ViaRandomKey/Reshuffle/GroupByKey/ReadStream+AvroIO.Write/AvroIO.TypedWrite/Write/GatherTempFileResults/Reshuffle.ViaRandomKey/Reshuffle/GroupByKey/MergeBuckets+AvroIO.Write/AvroIO.TypedWrite/Write/GatherTempFileResults/Reshuffle.ViaRandomKey/Reshuffle/ExpandIterable+A…
2019-10-22 02:29:35.188 BST  Executing operation AvroIO.Write/AvroIO.TypedWrite/Write/WriteShardedBundlesToTempFiles/GroupIntoShards/ReadStream+AvroIO.Write/AvroIO.TypedWrite/Write/WriteShardedBundlesToTempFiles/GroupIntoShards/MergeBuckets+AvroIO.Write/AvroIO.TypedWrite/Write/WriteShardedBundlesToTempFiles/WriteShardsIntoTempFiles+AvroIO.Write/AvroIO.TypedWrite/Write/GatherTempFileRe…
2019-10-22 02:29:35.253 BST  Executing operation AvroIO.Write/AvroIO.TypedWrite/Write/GatherTempFileResults/Reshuffle/GroupByKey/ReadStream+AvroIO.Write/AvroIO.TypedWrite/Write/GatherTempFileResults/Reshuffle/GroupByKey/MergeBuckets+AvroIO.Write/AvroIO.TypedWrite/Write/GatherTempFileResults/Reshuffle/ExpandIterable+AvroIO.Write/AvroIO.TypedWrite/Write/GatherTempFileResults/Drop key/Va…
2019-10-22 02:29:35.272 BST  Executing operation PubsubIO.Read/PubsubUnboundedSource+PubsubIO.Read/MapElements/Map+BigQueryIO.Write/PrepareWrite/ParDo(Anonymous)+Apply 1 Min Window/Window.Assign+Get Sale Lines/Map+BigQueryIO.Write/StreamingInserts/CreateTables/ParDo(CreateTables)+BigQueryIO.Write/StreamingInserts/StreamingWriteTables/ShardTableWrites+BigQueryIO.Write/StreamingInserts/…
2019-10-22 02:29:35.297 BST  Executing operation Count Store Sales/GroupByKey/ReadStream+Count Store Sales/Combine.GroupedValues+Create Data Point/Map+PubsubIO.Write/MapElements/Map+PubsubIO.Write/PubsubUnboundedSink
2019-10-22 02:29:35.322 BST  Executing operation BigQueryIO.Write/StreamingInserts/StreamingWriteTables/Reshuffle/GroupByKey/ReadStream+BigQueryIO.Write/StreamingInserts/StreamingWriteTables/Reshuffle/GroupByKey/MergeBuckets+BigQueryIO.Write/StreamingInserts/StreamingWriteTables/Reshuffle/ExpandIterable+BigQueryIO.Write/StreamingInserts/StreamingWriteTables/GlobalWindow/Window.Assign+…
2019-10-22 02:29:36.855 BST  Workers have started successfully.
```

Dataflow features

Centralized Logging

- Single searchable logging via GCP logging

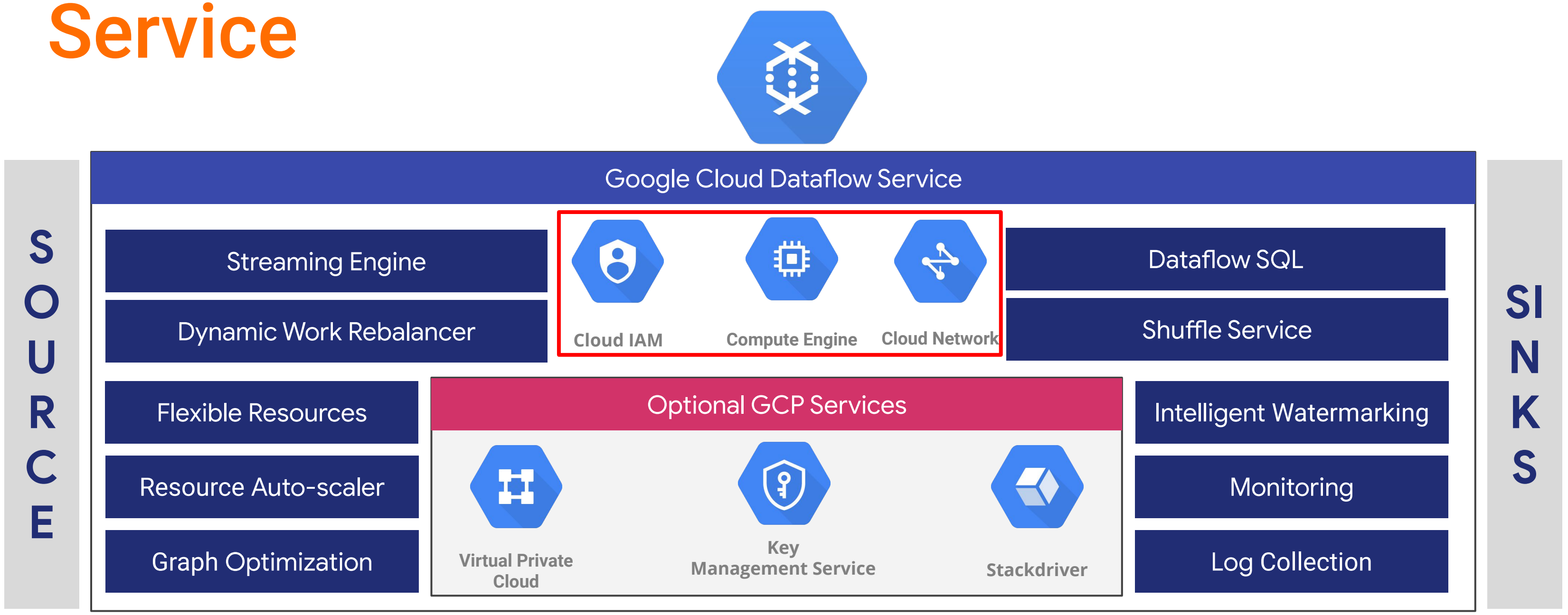




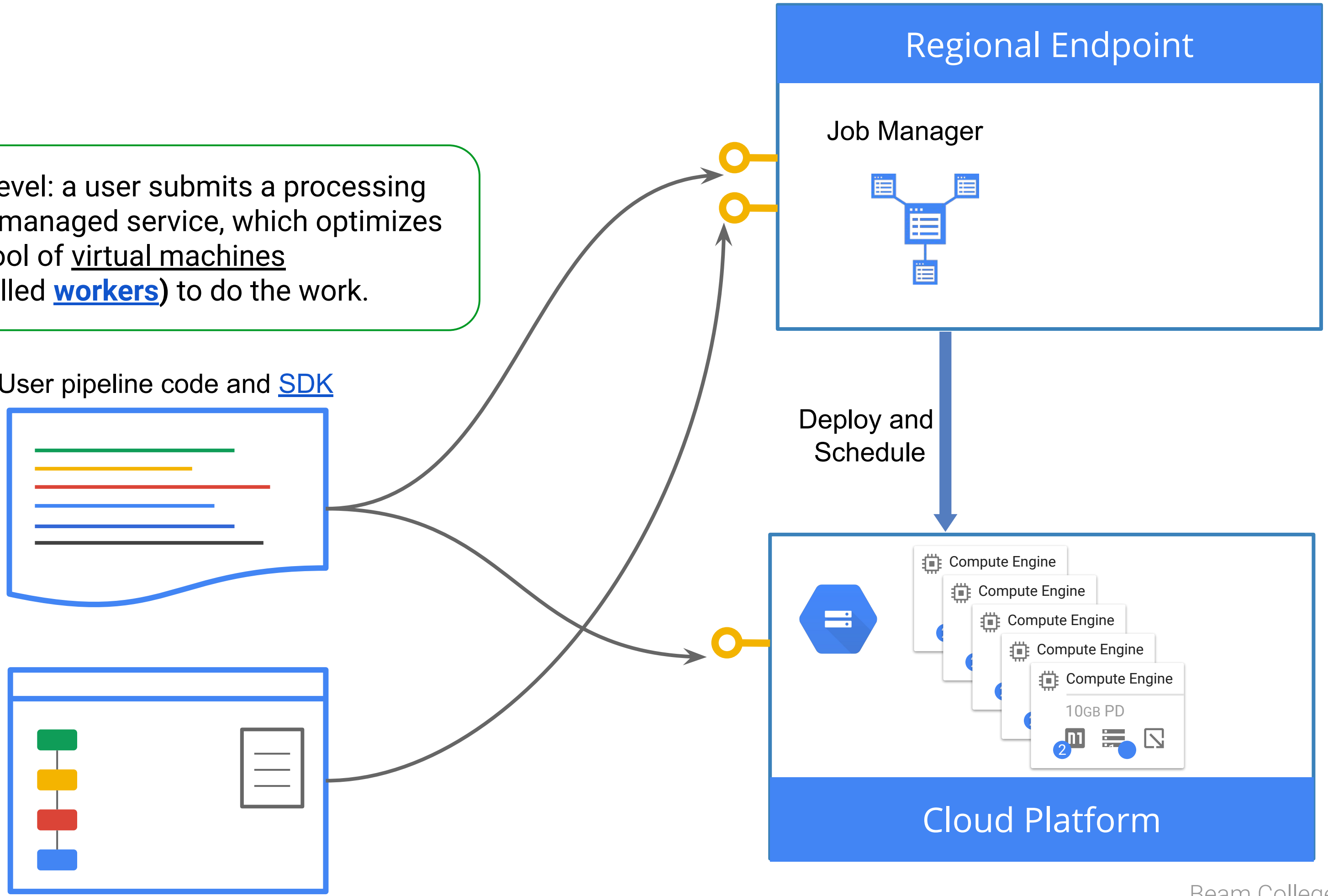
Showing logs from all time (BST)

```
2019-04-30 05:41:45.998 BST [pool-3-thread-7] INFO org.apache.beam.runners.dataflow.worker.DataflowBatchWorkerHarness - Finished processing stage s11 with…
2019-04-30 05:41:46.035 BST [pool-3-thread-1] INFO org.apache.beam.runners.dataflow.worker.WorkItemStatusClient - Success processing work item bigdatapivo…
2019-04-30 05:41:46.035 BST [pool-3-thread-1] INFO org.apache.beam.runners.dataflow.worker.DataflowBatchWorkerHarness - Finished processing stage s11 with…
2019-04-30 05:41:46.314 BST [pool-3-thread-6] INFO org.apache.beam.runners.dataflow.worker.DataflowBatchWorkerHarness - Starting MapTask stage s13
2019-04-30 05:41:46.314 BST [pool-3-thread-1] INFO org.apache.beam.runners.dataflow.worker.DataflowBatchWorkerHarness - Starting MapTask stage s13
2019-04-30 05:41:46.315 BST [pool-3-thread-8] INFO org.apache.beam.runners.dataflow.worker.DataflowBatchWorkerHarness - Starting MapTask stage s13
2019-04-30 05:41:46.345 BST [pool-3-thread-1] INFO org.apache.beam.runners.dataflow.worker.WorkItemStatusClient - Success processing work item bigdatapivo…
2019-04-30 05:41:46.345 BST [pool-3-thread-1] INFO org.apache.beam.runners.dataflow.worker.DataflowBatchWorkerHarness - Finished processing stage s13 with…
2019-04-30 05:41:46.346 BST [pool-3-thread-8] INFO org.apache.beam.runners.dataflow.worker.WorkItemStatusClient - Success processing work item bigdatapivo…
2019-04-30 05:41:46.346 BST [pool-3-thread-8] INFO org.apache.beam.runners.dataflow.worker.DataflowBatchWorkerHarness - Finished processing stage s13 with…
2019-04-30 05:41:46.368 BST [pool-3-thread-6] INFO org.apache.beam.runners.dataflow.worker.WorkItemStatusClient - Success processing work item bigdatapivo…
2019-04-30 05:41:46.368 BST [pool-3-thread-6] INFO org.apache.beam.runners.dataflow.worker.DataflowBatchWorkerHarness - Finished processing stage s13 with…
2019-04-30 05:41:46.639 BST [pool-3-thread-1] INFO org.apache.beam.runners.dataflow.worker.DataflowBatchWorkerHarness - Starting MapTask stage s15
2019-04-30 05:41:46.640 BST [pool-3-thread-5] INFO org.apache.beam.runners.dataflow.worker.DataflowBatchWorkerHarness - Starting MapTask stage s15
2019-04-30 05:41:46.642 BST [pool-3-thread-4] INFO org.apache.beam.runners.dataflow.worker.DataflowBatchWorkerHarness - Starting MapTask stage s15
2019-04-30 05:41:46.703 BST [pool-3-thread-4] INFO org.apache.beam.runners.dataflow.worker.WorkItemStatusClient - Success processing work item bigdatapivo…
2019-04-30 05:41:46.703 BST [pool-3-thread-4] INFO org.apache.beam.runners.dataflow.worker.DataflowBatchWorkerHarness - Finished processing stage s15 with…
2019-04-30 05:41:46.705 BST [pool-3-thread-1] INFO org.apache.beam.runners.dataflow.worker.WorkItemStatusClient - Success processing work item bigdatapivo…
2019-04-30 05:41:46.705 BST [pool-3-thread-1] INFO org.apache.beam.runners.dataflow.worker.DataflowBatchWorkerHarness - Finished processing stage s15 with…
2019-04-30 05:41:46.764 BST [pool-3-thread-5] INFO org.apache.beam.runners.dataflow.worker.WorkItemStatusClient - Success processing work item bigdatapivo…
2019-04-30 05:41:46.764 BST [pool-3-thread-5] INFO org.apache.beam.runners.dataflow.worker.DataflowBatchWorkerHarness - Finished processing stage s15 with…
```

Google Cloud Dataflow Service



At a very high level: a user submits a processing pipeline to our managed service, which optimizes it and runs a pool of virtual machines (sometimes called **workers**) to do the work.



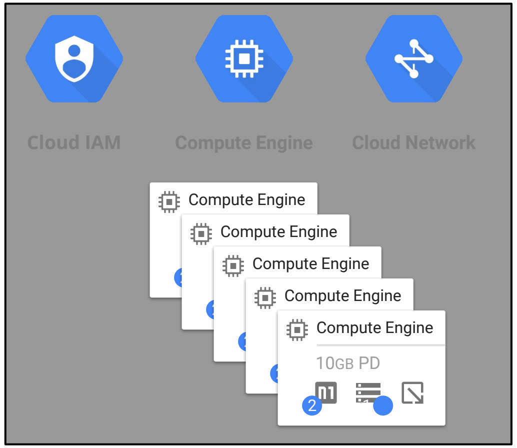
Dataflow & Compute Engine

Region endpoint

- Deploys and controls Dataflow workers and stores Job Metadata
- Region is **us-central1** by default, unless explicitly set using the region parameter

Zone

- Defines the locations of the Dataflow workers
- Defaults to a zone in the region selected based on available zone capacity. It can be overridden using the zone parameter.



The zone does not need to be in the same region as the endpoint. Reasons you may want to do this include:

- Security and Compliance
- Data locality
- Resilience and geographic separation

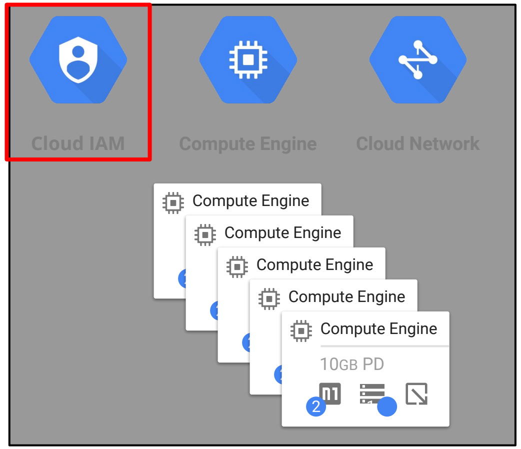
Caution: If you override the zone and the zone is in a different region than the regional endpoint, there may be negative impact on performance, network traffic, and network latency.

Dataflow & Compute Engine

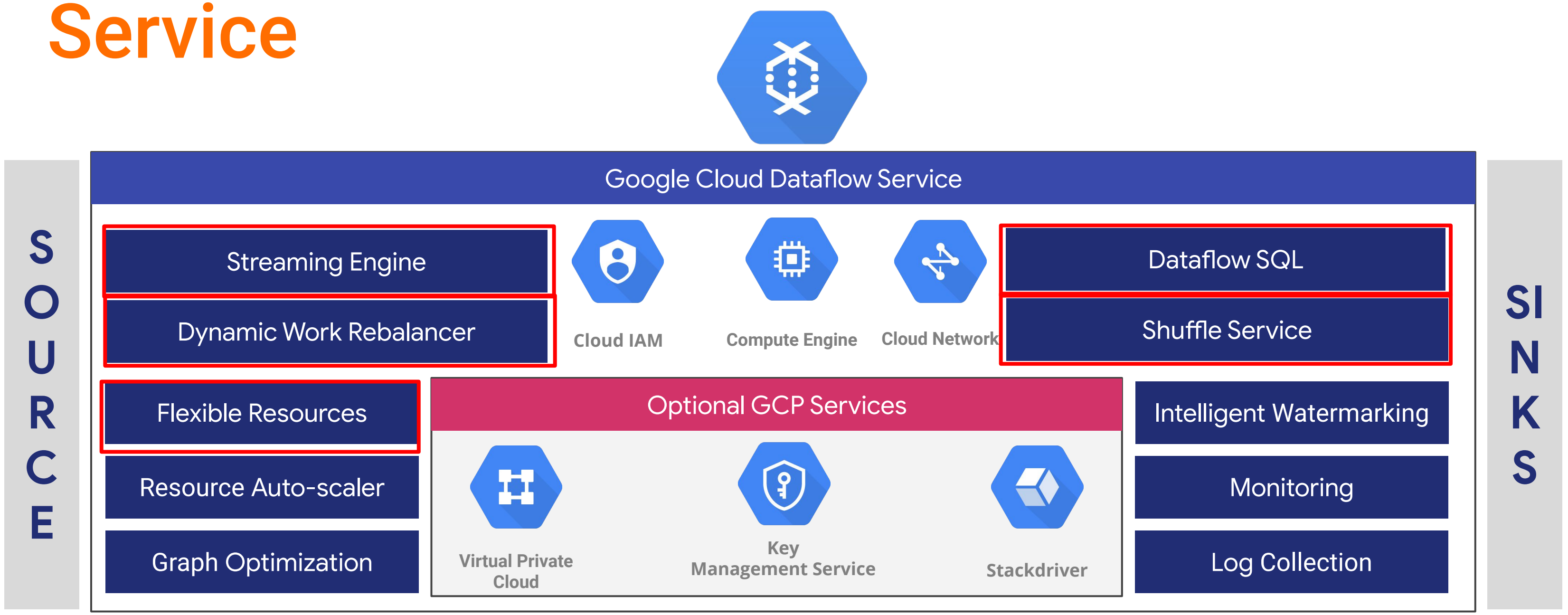
Identity Access Management

There are a minimum of 2 service accounts used by the Dataflow service

- Dataflow Service Account (`service-<project-number>@dataflow-service-producer-prod.iam.gserviceaccount.com`)
  - Used for worker creation, monitoring etc...
- Controller Service Account
  - `<project-number>-compute@developer.gserviceaccount.com`
  - Used by the workers to access resources needed by the pipeline, for example files on a Google Cloud Storage Bucket
  - Can be overridden using **--serviceAccount**



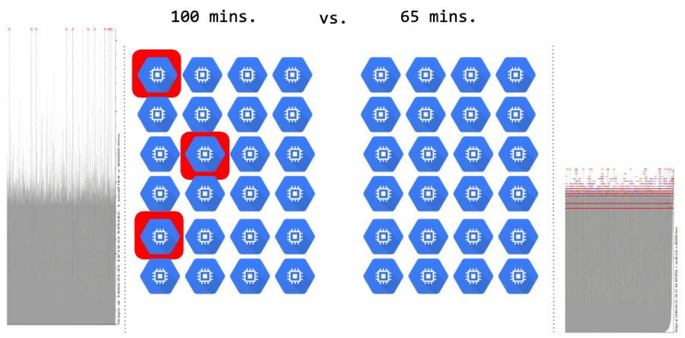
Google Cloud Dataflow Service



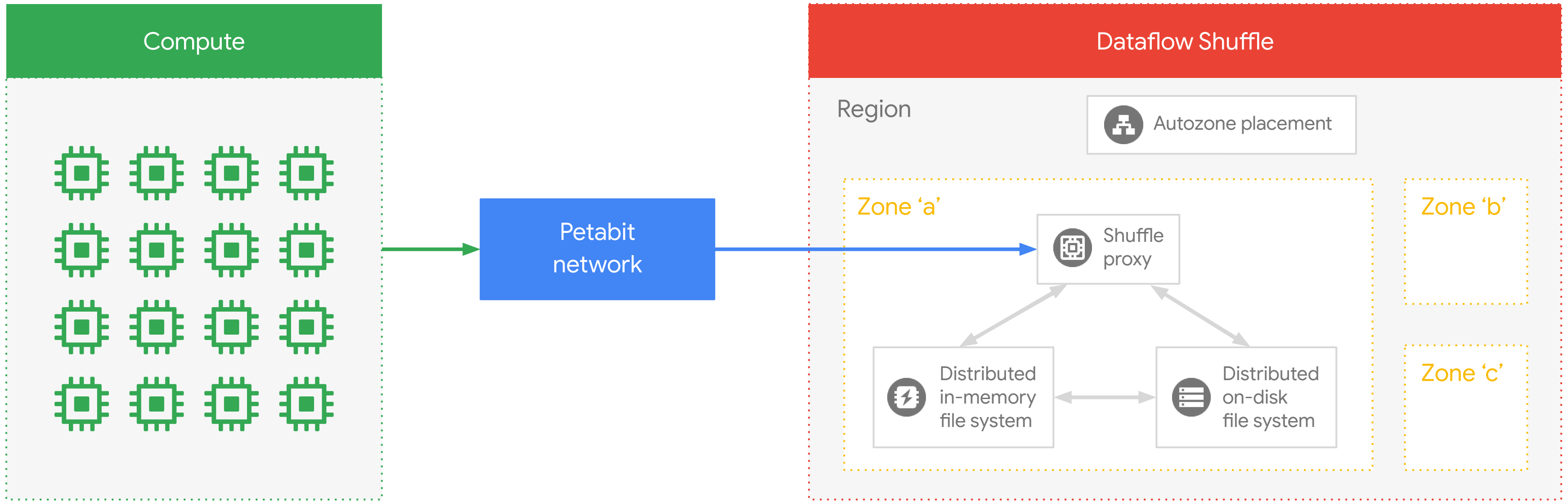
Dataflow features

Batch Dynamic Work Redistribution

- Redistribute hot keys for more even workload distribution.
- Fully automated



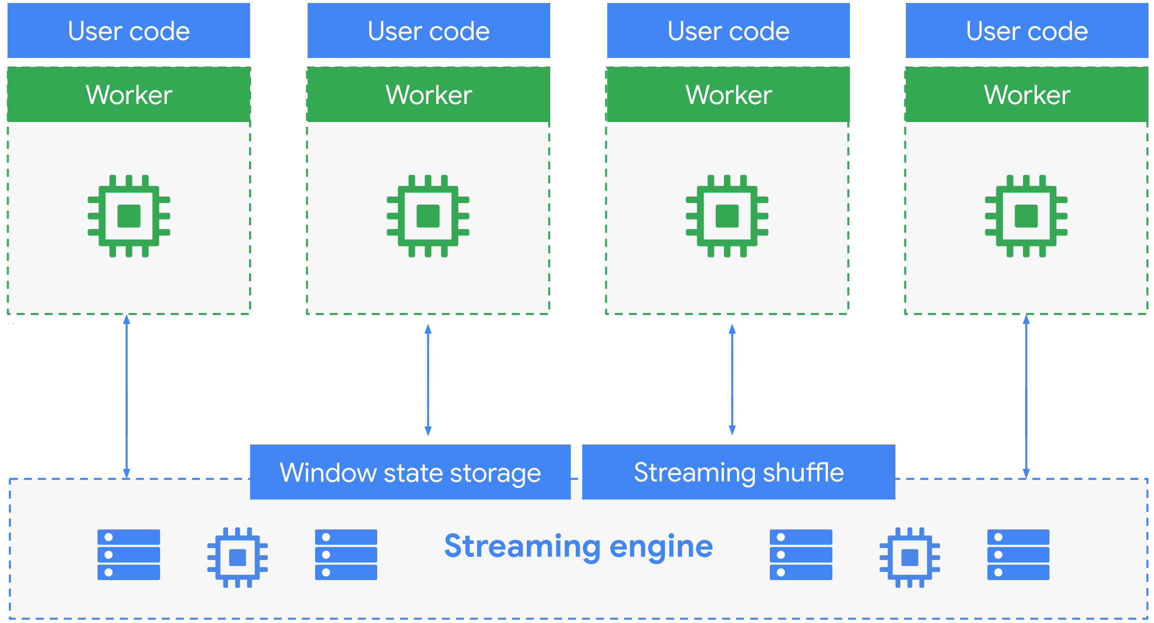
Dataflow Shuffle - Batch



Dataflow Streaming Engine

Benefits

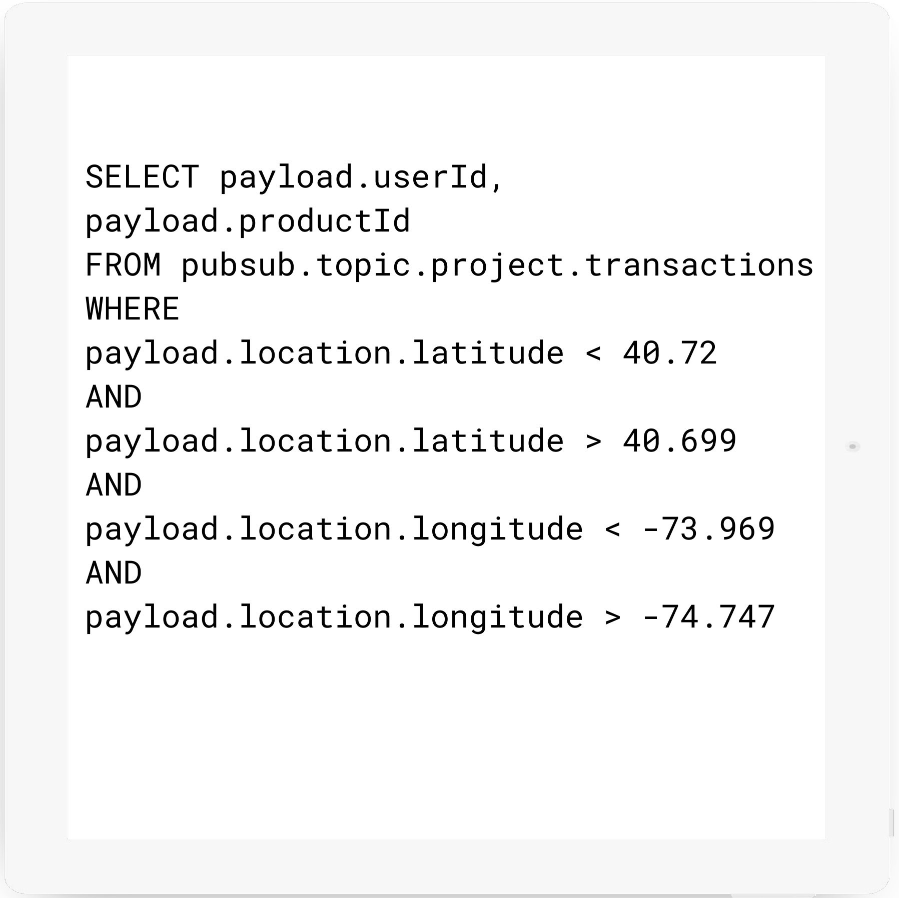
- ✓ Smoother autoscaling
- ✓ Better supportability
- ✓ Less worker resources



Dataflow SQL UI

No coding required

- Write SQL in BigQuery UI
- Use Schemas form Data Catalog
- Submit Dataflow jobs

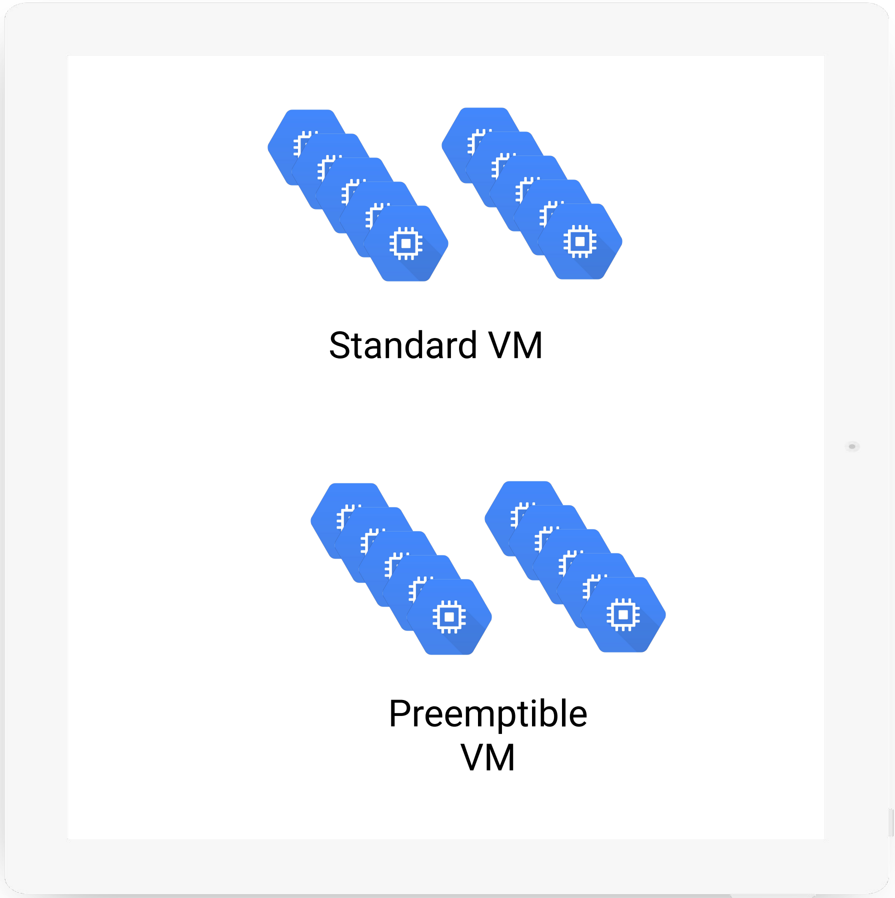


```
SELECT payload.userId,
payload.productId
FROM pubsub.topic.project.transactions
WHERE
payload.location.latitude < 40.72
AND
payload.location.latitude > 40.699
AND
payload.location.longitude < -73.969
AND
payload.location.longitude > -74.747
```

Dataflow features

Flexible Resource Scheduling

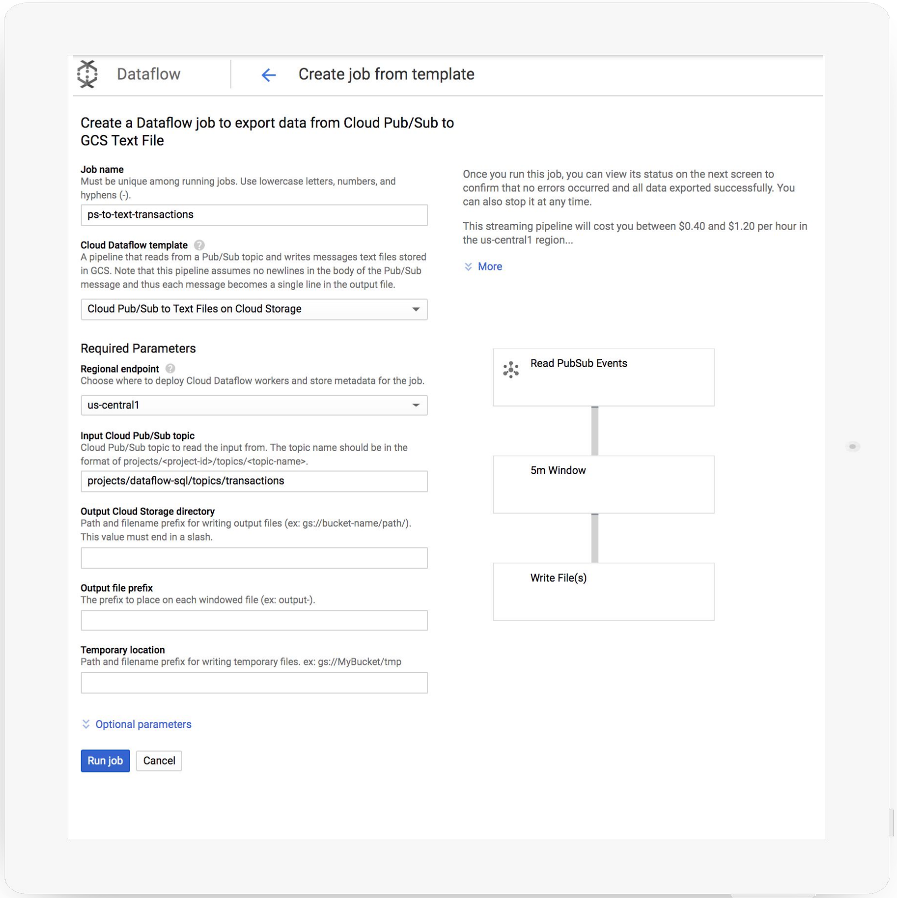
- FlexRS reduces batch processing costs by using advanced scheduling techniques, the Cloud Dataflow Shuffle service, and a combination of preemptible virtual machine (VM) instances and regular VMs.
- Jobs with FlexRS use service-based Cloud Dataflow Shuffle for joining and grouping.



Dataflow Templates

No coding required

- Select one of 20+ Google-provided templates or use your own
- Popular ETL sources and sinks
- Streaming and Batch modes
- Launch from GCS or Pub/Sub browsers



Dataflow ← Create job from template

Create a Dataflow job to export data from Cloud Pub/Sub to GCS Text File

Job name
Must be unique among running jobs. Use lowercase letters, numbers, and hyphens (-).
ps-to-text-transactions

Cloud Dataflow template
A pipeline that reads from a Pub/Sub topic and writes messages text files stored in GCS. Note that this pipeline assumes no newlines in the body of the Pub/Sub message and thus each message becomes a single line in the output file.
Cloud Pub/Sub to Text Files on Cloud Storage

Required Parameters

Regional endpoint
Choose where to deploy Cloud Dataflow workers and store metadata for the job.
us-central1

Input Cloud Pub/Sub topic
Cloud Pub/Sub topic to read the input from. The topic name should be in the format of projects/<project-id>/topics/<topic-name>.
projects/dataflow-sql/topics/transactions

Output Cloud Storage directory
Path and filename prefix for writing output files (ex: gs://bucket-name/path/). This value must end in a slash.

Output file prefix
The prefix to place on each windowed file (ex: output-).

Temporary location
Path and filename prefix for writing temporary files. ex: gs://MyBucket/tmp

Once you run this job, you can view its status on the next screen to confirm that no errors occurred and all data exported successfully. You can also stop it at any time.

This streaming pipeline will cost you between $0.40 and $1.20 per hour in the us-central1 region...

GPU Support

Attach Graphical Processing Units (GPU) to your Dataflow workers to accelerate ML model training, batch and Streaming predictions, and general data processing

What you can do with it

- Select from a range of GPU types (NVIDIA K80, P100, P4, T4, and V100s) for your job
- Accelerate ML workloads (preprocessing, feature engineering, ML inference)

Dataflow Prime

Ingest and distribute data reliably with Serverless and OSS systems

amazon S3

HDFS

Dataflow Serverless

Serverless Auto Tuning Infrastructure

Serverless Smart Diagnostics

Simplified Billing

“Streaming ML” for real time insights

Unified Batch and Streaming

Open, intelligent and flexible platform

Store and analyze at scale with serverless and OSS systems

Governance, Security, Lineage and Workflow Management

Summary

1. Architecture
2. DAG optimization
3. Shuffle Service
4. Streaming Engine
5. Monitoring / Logging
6. Flexible Resource Scheduling
7. Out of the box Templates
8. SQL UI
9. Dataflow Prime