



Modern ETL Tools for Cloud and Big Data

Ken Beutler, Principal Product Manager, Progress
Michael Rainey, Technical Advisor, Gluent Inc.



Agenda

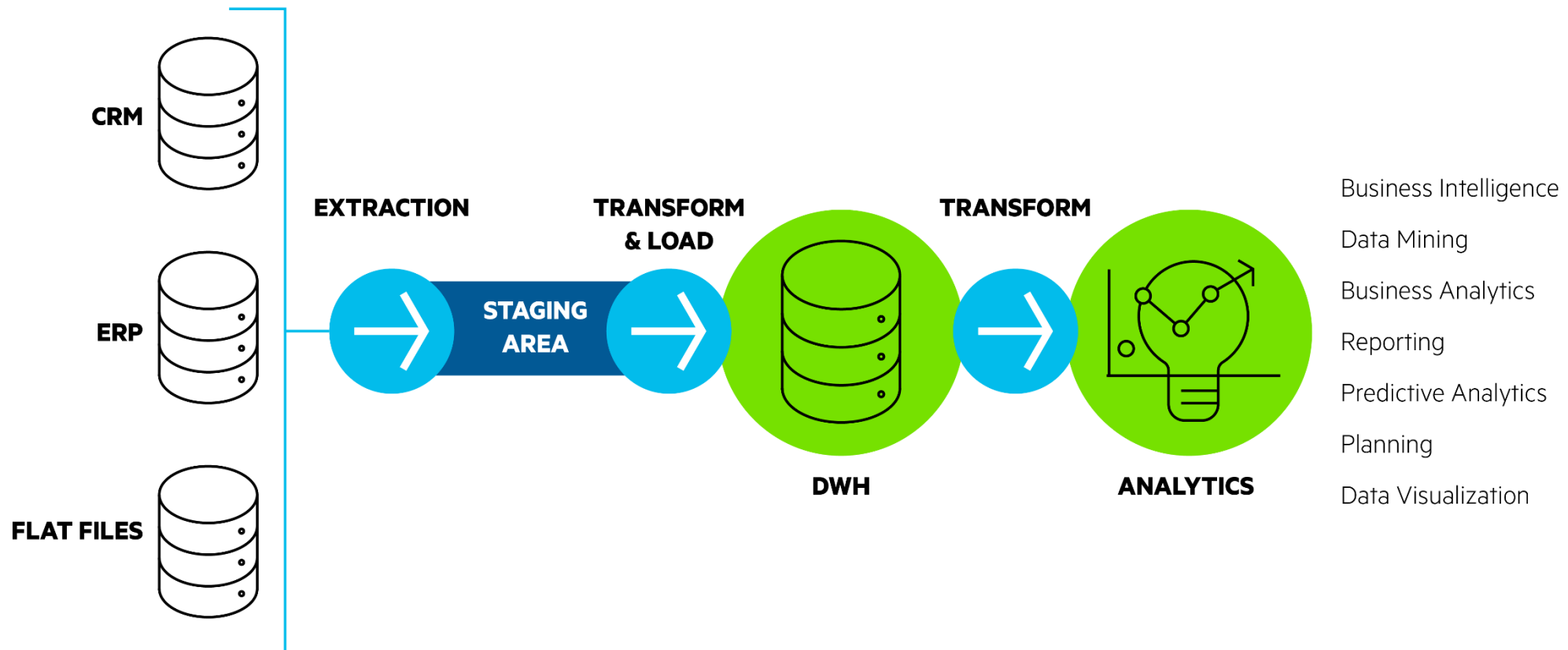
- Landscape
- Cloud ETL Tools
- Big Data ETL Tools
- Best Practices
- QnA



Landscape



Data Warehousing Has Transformed Enterprise BI and Analytics



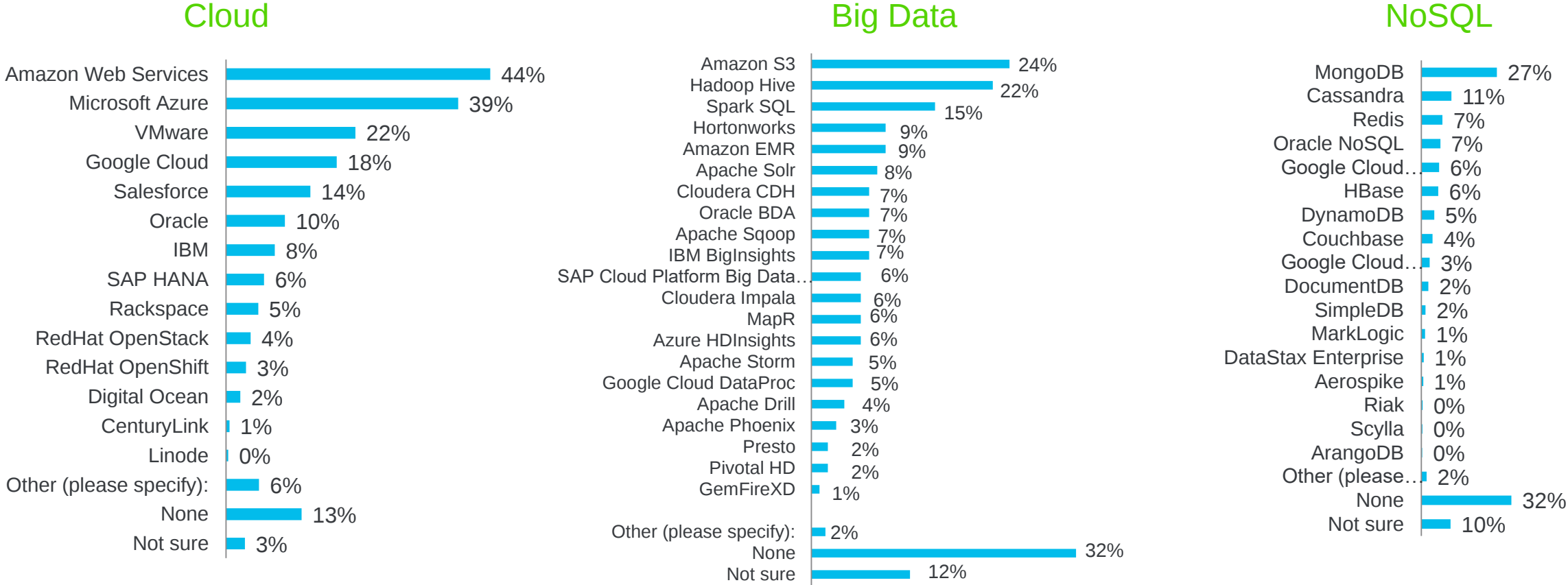
ETL Tools Are At The Heart Of EDW



Rise Of The Modern EDW Solutions



Enterprises Are Increasingly Adopting Cloud and Big Data



Source: Progress DataDirect's Data Connectivity Outlook Survey 2018



Cloud ETL Tools



Poll Question 1

- Which cloud provider are you currently using / plan to use?
- AWS
- Google Cloud Platform
- Azure
- Other
- None



AWS Glue

Serverless ETL

AWS Glue

■ Components:

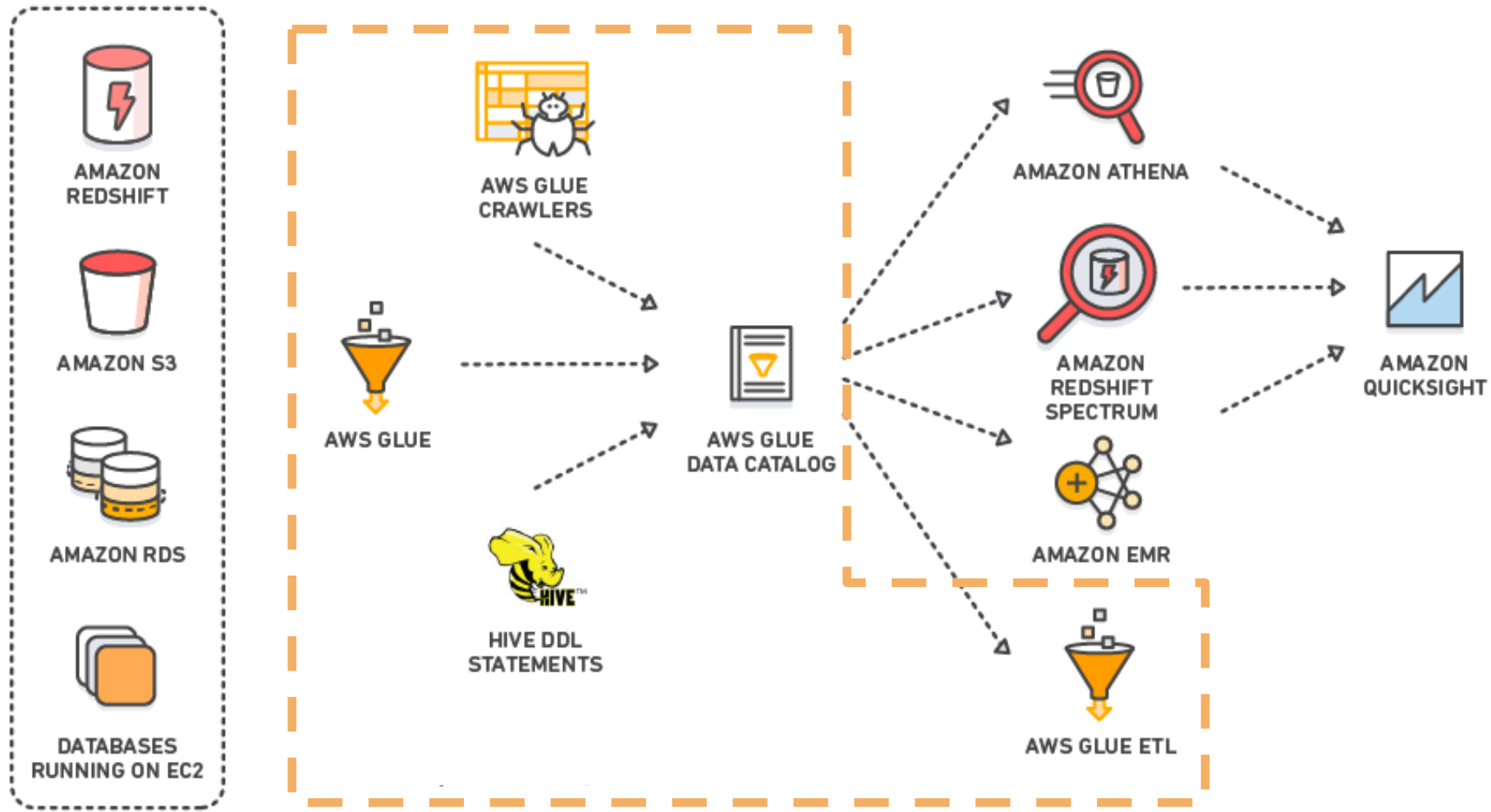
- Data Catalog
- Crawlers
- ETL jobs/scripts
- Job scheduler



■ Useful for...

- ...running serverless queries against S3 buckets and relational data
- ...creating event-driven ETL pipelines
- ...automatically discovering and cataloging your enterprise data assets

AWS Glue architecture



Source: <https://docs.aws.amazon.com/athena/latest/ug/glue-athena.html>

© 2018 Progress Software Corporation and/or its subsidiaries or affiliates. All rights reserved.

Data catalog - the central component

- Can act as metadata repository for other Amazon services

- Tables - Added to “databases” using the wizard or a crawler

☐ Name ▼	Database ▲	Location ▼	Classification
☐ channels	int12102_sh	s3://gluent.backup/user/gluent/b...	orc
☐ countries	int12102_sh	s3://gluent.backup/user/gluent/b...	parquet
☐ customers	int12102_sh	s3://gluent.backup/user/gluent/b...	parquet
☐ gl_chars	int12102_sh	s3://gluent.backup/user/gluent/b...	parquet

- Data sources: Amazon S3, Redshift, Aurora, Oracle, PostgreSQL, MySQL, MariaDB, MS SQL Server, JDBC, DynamoDB
- **Crawlers** connect to one or more data stores, determine the data structures, and write tables into the Data Catalog

Jobs

ETL

- PySpark or Scala scripts, generated by AWS Glue
- Visual dataflow can be generated, but not used for development
- Execute ETL using the job scheduler, events, or manually invoke
- Built-in transforms used to process data

- **ApplyMapping**

- Maps source and target columns

- **Filter**

- Load new DynamicFrame based on filtered records

- **Join**

- Joins two DynamicFrames

- **SelectFields**

- Output selected fields to new DynamicFrame

- **SplitRows**

- Split rows into two new DynamicFrames based on a predicate

- **SplitFields**

- Split fields into two new DynamicFrames

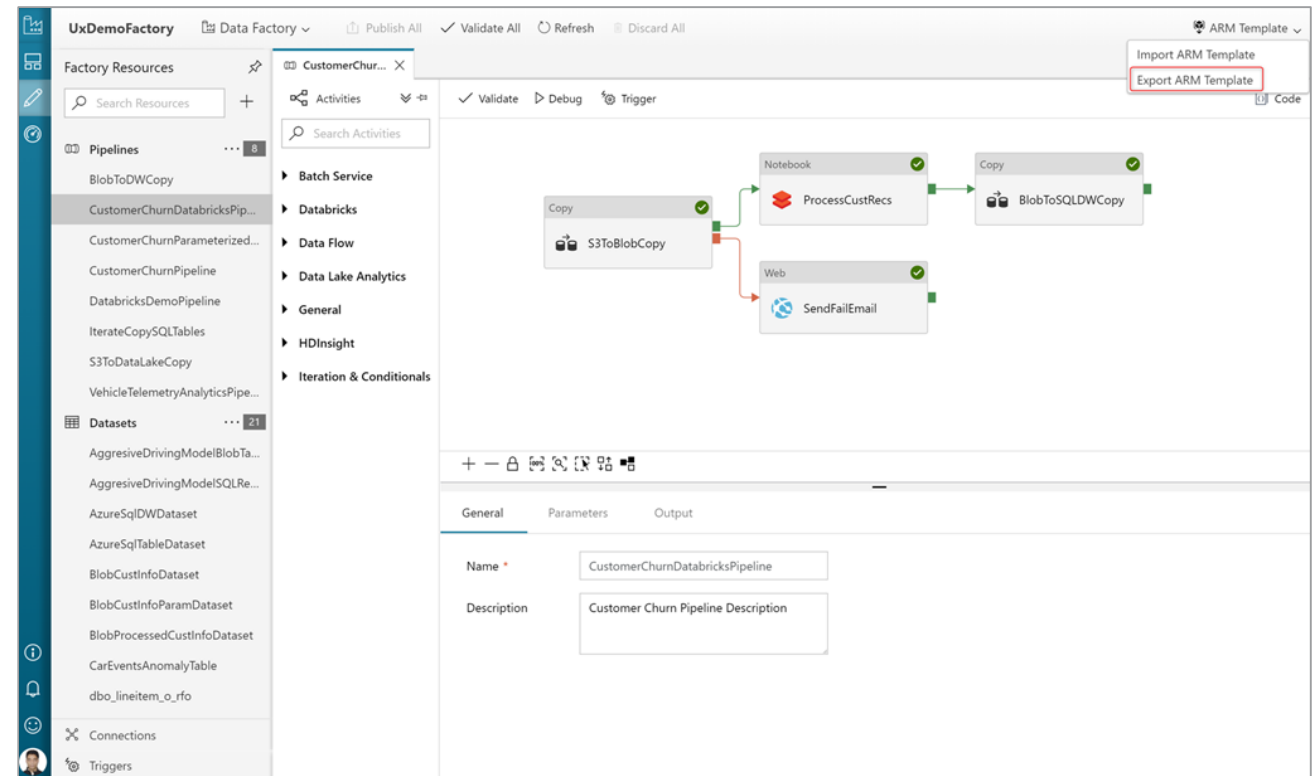


Azure Data Factory

Visual Cloud ETL

Azure Data Factory

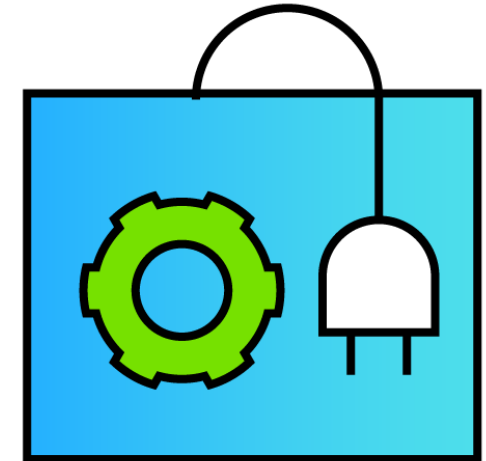
- Build data **pipelines** using a visual ETL user interface
 - Visual Studio Team Services (VSTS) Git integration for collaboration, source control, and versioning
- Drag, drop, link **activities**
 - **Copy Data:** Source to Target
 - **Transform:** Spark, Hive, Pig, streaming on HDInsight, Stored Procedures, ML Batch Execution, etc.
 - **Control flow:** If-then-else, For-each, Lookup, etc.



Source: <https://azure.microsoft.com/en-us/blog/continuous-integration-and-deployment-using-data-factory/>

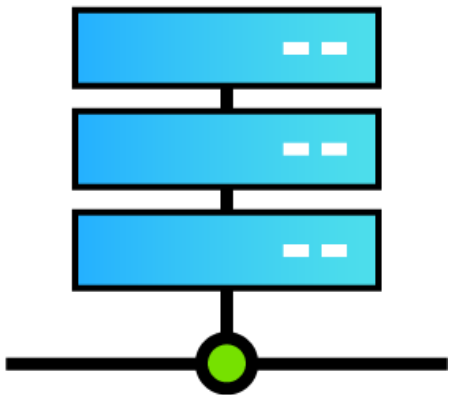
Linked services

- Allow connection to many data sources
 - Source:
 - Azure DBs, Azure Blob Storage, Azure File Storage, Oracle, SQL Server, SAP HANA, Sybase, DB2, Impala, Drill, MySQL, Google BigQuery, Cassandra, Amazon S3, HDFS, Salesforce, etc.
 - Target:
 - Azure DBs, Azure Blob Storage, Azure File Storage, Oracle, SQL Server, SAP HANA, File System, Generic ODBC, Salesforce, etc.



Integration runtime environment

- Compute environment required for Azure Data Factory
 - Fully managed in Azure cloud infrastructure
 - Automatic scaling, high availability, etc.
 - Allows execution of SSIS packages

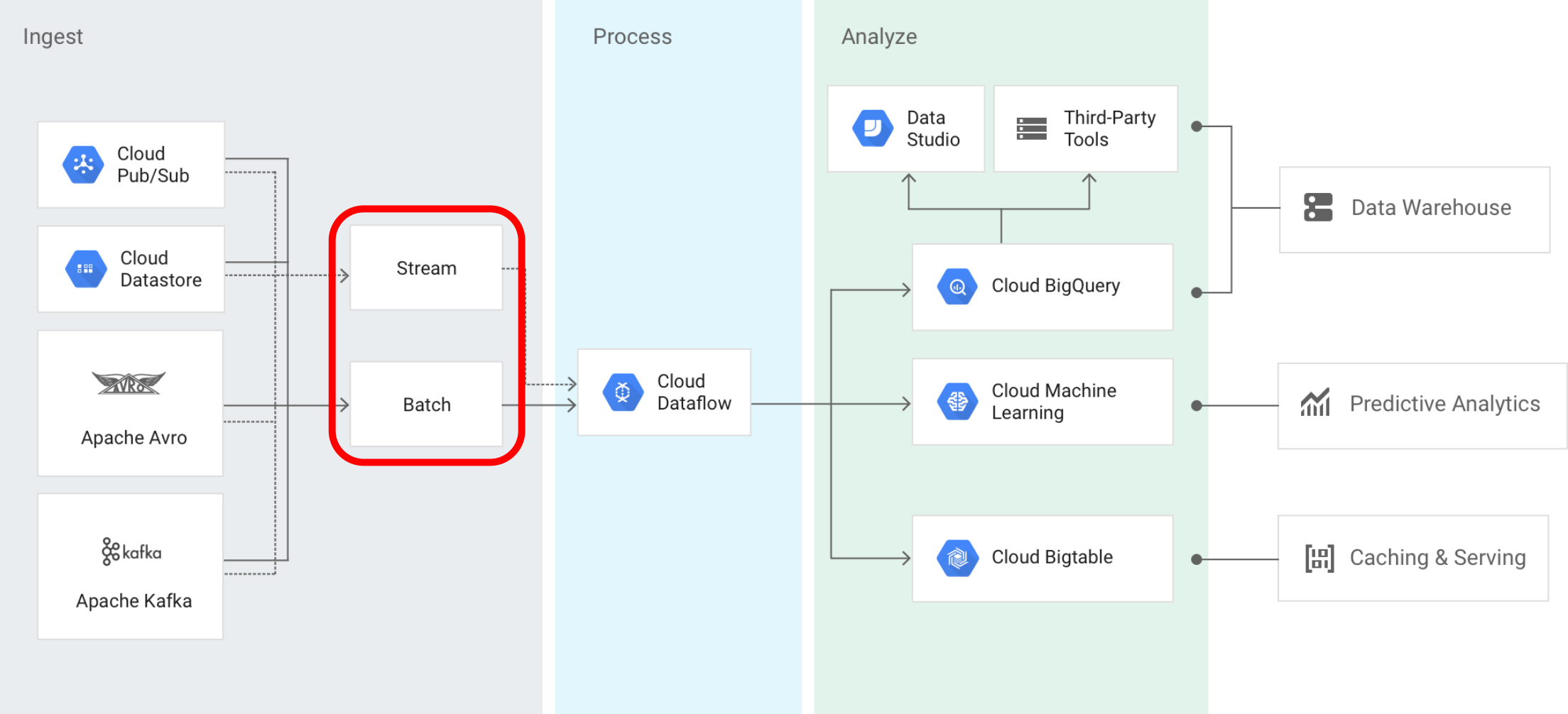


- Can be hosted on a local private network
 - Organizations can control, and manage, compute resources

Google Cloud Dataflow

Unified Programming Model for Data Processing

Google Cloud Dataflow



Source: <https://cloud.google.com/dataflow/>

Google Cloud Dataflow overview

- Unified programming model for batch (historical) and streaming (real-time) data pipelines *and* distributed compute platform
 - Reuse code across both batch and streaming pipelines
 - Java or Python based
- Programming model an open source project - Apache Beam (<https://beam.apache.org/>)
 - Runs on multiple different distributed processing back-ends: Spark, Flink, Cloud Dataflow platforms
- Fully managed service
 - Automated resource management and scale-out



Source: <https://cloud.google.com/dataflow/>

Dataflow I/O transforms

Language	File-based	Messaging	Database
Java	Beam Java supports Apache HDFS, Amazon S3, Google Cloud Storage, and local filesystems FileIO, AvroIO, TextIO, TFRecordIO, XmlIO, TikaIO, ParquetIO Many different filesystems and cloud object stores supported	Amazon Kinesis AMQP Apache Kafka Google Cloud Pub/Sub JMS MQTT	Apache Cassandra Apache Hadoop InputFormat Apache HBase Apache Hive (HCatalog) Apache Solr JDBC ElasticSearch Google BigQuery Google Cloud BigTable Google Cloud Datastore Google Cloud Spanner MongoDB Redis JDBC allows custom connection to additional databases Other I/O transforms are already under development
Python	Beam Python supports Google Cloud Storage and local filesystems avroio, textio, tfrecordio, vc fio	Google Cloud Pub/Sub	Google BigQuery Google Cloud Datastore

Cloud ETL tools compared

Batch ETL
Streaming
User Interface
Compute data platform
Cross-platform support
Custom connector support
Metadata catalog
Monitoring tools available
Fully managed



Big Data ETL Tools



Poll Question 2

- Which of the following tools are you currently using / plan to use?
- Apache Sqoop
- Apache Nifi
- Apache Flume
- Other
- None



Apache Sqoop

Batch Data Transfer

Sqoop Overview



- Sqoop is a tool designed to transfer data between Hadoop and RDBMS or mainframes (bi-directional)
- Open source (Apache) command line tool
- Batch/Bulk transfer of structured data
- Requires Hadoop
 - Leverages MapReduce to provide parallel execution and data partitioning

Sqoop Details



Importing Data

- Bulk move data from RDBMS or mainframe to Hadoop
 - Persist in HDFS, Hive, HBase or Accumulo
- Uses JDBC* for connectivity
- Able to perform column and row filtering or free-form queries
 - Must use MapReduce, Pig or other scripting language

* Very limited number of databases are supported and version specific

Exporting Data

- Exports a set of files in HDFS to RDBMS
- Uses JDBC* for connectivity
- Leverages either update or insert mode or incremental imports
- Limited transformation support – column filtering
- Constraints:
 - Table must exist
 - Operation is not atomic

Sqoop Details



- Sqoop jobs can be created and saved to be used as templates
- Sqoop jobs can be scheduled using Oozie
- sqoop-merge can be used to combine two datasets into one and overwrite values of an older dataset
- Plugin architecture

```
$ sqoop import --connect jdbc:mysql://localhost/userdb --username root --table emp --m 1
```

Apache Flume

Log/Event collection, aggregation and transfer



Flume Overview

- Flume is a distributed, reliable and available service for collecting, aggregating and moving large amounts of streaming data into a backing store
- Open source (Apache) command line tool
- Transfer of high throughput, low latency, structured data
 - Application logs, sensor, machine, geo-location and social data
- Does not require Hadoop but has many integrations with services that can leverage Hadoop effectively

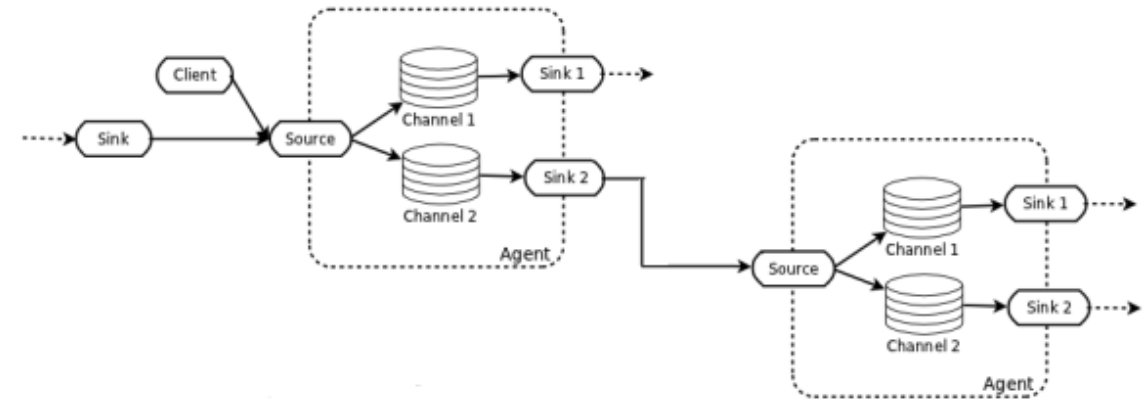
Flume Details



Ingesting data

- Event – A singular unit of data that is transported
- Source – The entity through which data enters into Flume. Sources either poll for data or passively wait for data
- Sink – Delivers data to the destination. A variety of sinks allow data to be streamed to a variety of destinations
- Channel – The conduit between the source and the sink. Sources ingest events into the channel and sinks drain the channel
- Agent – A collection of sources, sinks and channels running in a JVM

Data Flow Model





Apache Nifi

Process, track and distribute data

Nifi Overview



- Platform tool built to automate and manage the flow of data between systems
 - Visual Command and Control
 - Data Provenance
 - Data Prioritization
 - Data Buffering/Back-Pressure
 - Control Latency vs. Throughput
 - Security
 - Scale Out Clustering
 - Extensibility
- When to Use Nifi?
- When to Not Use Nifi?

NiFi Summary

PROCESSORS INPUT PORTS OUTPUT PORTS REMOTE PROCESS GROUPS CONNECTIONS PROCESS GROUPS

Displaying 60 of 60

Filter by name View: Single node Cluster

	Name ^	Type	Run Status	In / Size 5 min	Read / Write 5 min	Out / Size 5 min	Tasks / Time 5 min	
❗	Base64EncodeCo...	Base64EncodeCo...	⚠ Invalid	0 (0 bytes)	0 bytes / 0 bytes	0 (0 bytes)	0 / 00:00:00.000	→ 🚚
❗	Capture Network ...	ExecuteProcess	▶ Running (5)	0 (0 bytes)	0 bytes / 245.97 MB	2,883 (245.97 MB)	2,911 / 00:24:59.988	→ 🚚
❗	Check if Dataset ...	IdentifyMimeType	▶ Running	0 (0 bytes)	0 bytes / 0 bytes	0 (0 bytes)	0 / 00:00:00.000	→ 🚚
❗	Decompress Data...	CompressContent	▶ Running	0 (0 bytes)	0 bytes / 0 bytes	0 (0 bytes)	0 / 00:00:00.000	→ 🚚
❗	Detect Interesting...	RouteOnAttribute	▶ Running (2)	11,741,700 (13 GB)	0 bytes / 0 bytes	0 (0 bytes)	11,741,700 / 00:09:...	→ 🚚
❗	Empty Contents	ReplaceText	▶ Running	0 (0 bytes)	0 bytes / 0 bytes	0 (0 bytes)	0 / 00:00:00.000	→ 🚚
❗	EvaluateJsonPath	EvaluateJsonPath	▶ Running	0 (0 bytes)	0 bytes / 0 bytes	0 (0 bytes)	0 / 00:00:00.000	→ 🚚
❗	EvaluateJsonPath	EvaluateJsonPath	▶ Running (4)	11,742,510 (13 GB)	13 GB / 0 bytes	11,742,510 (13 GB)	11,742,856 / 00:20:...	→ 🚚
❗	Extract File ID	UpdateAttribute	▶ Running	0 (0 bytes)	0 bytes / 0 bytes	0 (0 bytes)	0 / 00:00:00.000	→ 🚚
❗	Extract File ID	UpdateAttribute	▶ Running	0 (0 bytes)	0 bytes / 0 bytes	0 (0 bytes)	0 / 00:00:00.000	→ 🚚
❗	ExtractText	ExtractText	▶ Running	1,498,914 (244.45 ...)	244.45 MB / 0 bytes	1,496,995 (244.25 ...)	1,500,692 / 00:02:4...	→ 🚚

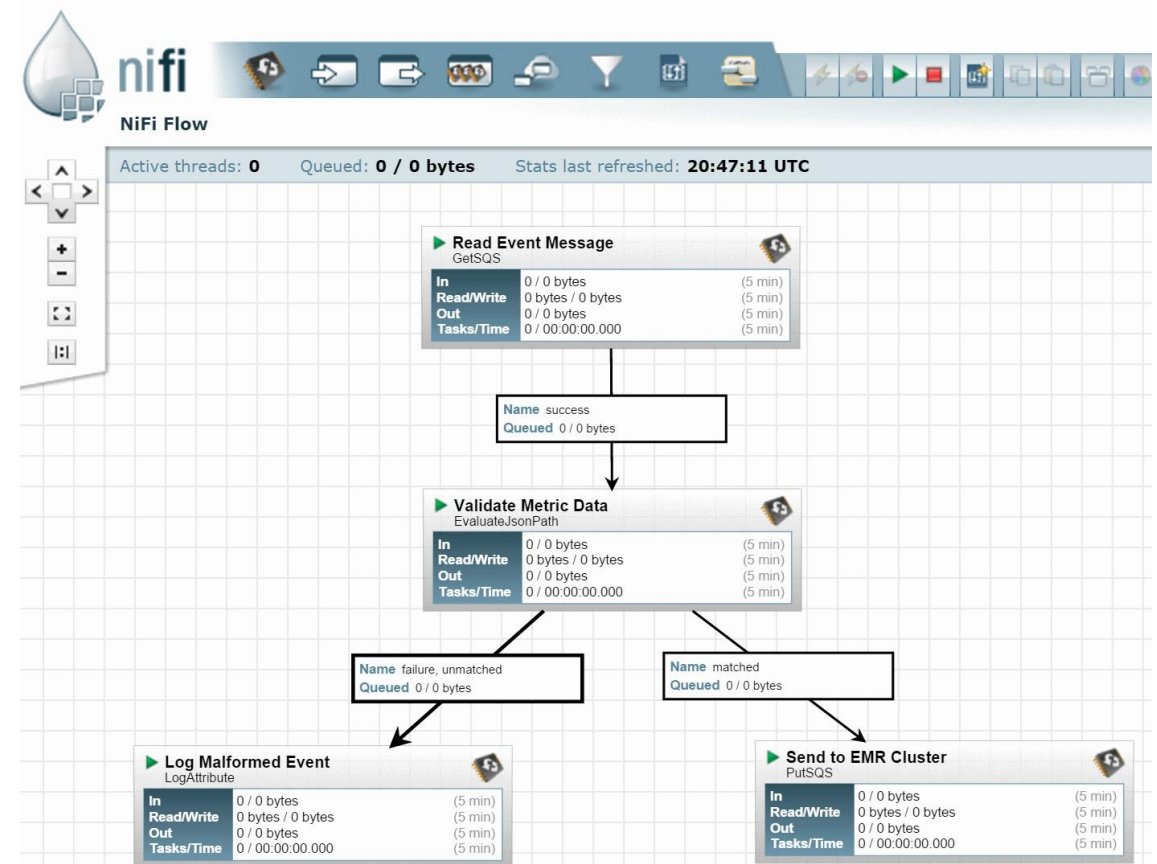
🔄 Last updated: 17:45:24 UTC [system diagnostics](#)

Nifi – Data Flows



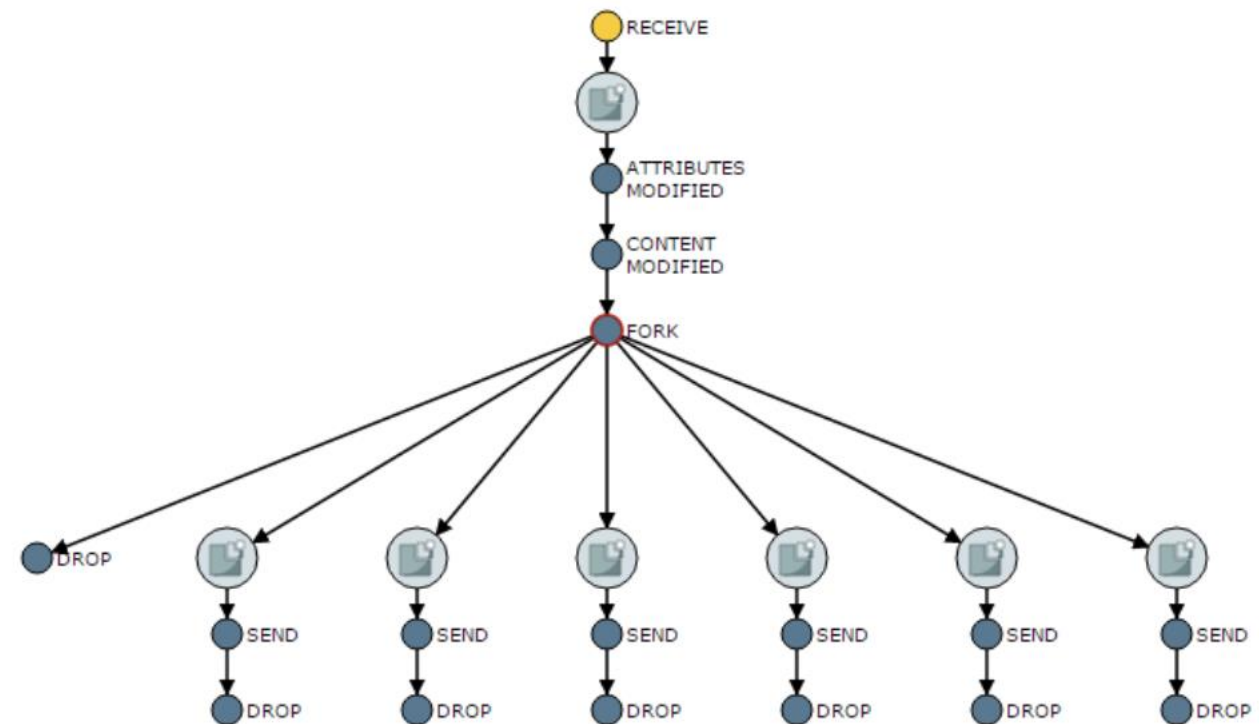
Components of a Data Flow

- FlowFile
- FlowFile Processor
- Connection
- Flow Controller
- Process Group



Nifi – Data Provenance

- Records the provenance of data as it flows through the system
- Searching, filtering and drill-in capabilities
- Events
 - receive, send, clone, drop, fork, route, modify, join...



Comparison of Big Data ETL Tools

ETL Tool	Purpose	Limitation(s)	Use Cases
Sqoop	• Command line tool for bi-directional bulk data movement	• OOTB connectivity does not support many versions or sources • Can put strain on data source as the MapReduce jobs partition the data	• Extract operational data from RDBMS for processing in Hadoop • Data archiving of historical or expired data to Hadoop
Flume	• Command line tool for collecting, aggregating and moving event-based data	• Message delivery is not guaranteed in some instances • Channel backing store can limit scalability	• Detect and report on web traffic anomalies in near real-time • Monitoring global network traffic for threat detection and fault tolerance
Nifi	• Command and control framework for modeling, monitoring and moving data between systems	• May have scalability issues with extreme volume/low latency data flows	• Everything above but greater flexibility in terms of developer and user experience

Poll Question 3

- What data sources are most challenging to connect to from your ETL tool?
- SaaS sources (Salesforce, Eloqua, Oracle Service Cloud, Google Analytics, etc.)
- Relational Sources (Oracle, SQL Server, IBM DB2, etc.)
- Data Warehouses (RedShift, Teradata, Sybase, Greenplum, etc.)
- Big Data Sources (Hadoop, MongoDB, Cassandra, etc.)
- Other



EDW/ETL Best Practices



EDW/ETL Best Practices



Simplify Access

- Self service
- Data breadth
- Tool training



Data Governance

- Create a data governance board
- Appoint data stewards
- Build a DQ firewall



Data Quality

- DQ assessment
- DQ measurement
- Incorporate DQ into functions/processes
- Let the business drive DQ



Data Modeling

- Schema on read and schema on write
- Normalized and de-normalized schemas
- Logical and physical data models

Resources

■ Tutorials:

- [AWS Glue + Salesforce JDBC Driver](#)
- [Google DataFlow + Salesforce into Google BigQuery](#)
- [Apache Sqoop: EXPORT to SQL Server from Hadoop](#)
- [Apache Nifi: Ingest Salesforce Data Incrementally into Hive](#)

■ *Progress DataDirect:*

- [JDBC](#) and [ODBC](#) Connectors
- [Why](#) DataDirect JDBC Drivers?

■ *Gluent:*

- *Integrate Enterprise Data with [Gluent Cloud Sync and AWS Glue](#)*



Q&A



