**Data Engineering Summary**

Here's a quick summary of all the services we've mentioned

* Amazon S3: Object Storage for your data
* VPC Endpoint Gateway: Privately access your S3 bucket without going through the public internet
* Kinesis Data Streams: real-time data streams, need capacity planning, real-time applications
* Kinesis Data Firehose: near real-time data ingestion to S3, Redshift, ElasticSearch, Splunk
* Kinesis Data Analytics: SQL transformations on streaming data
* Kinesis Video Streams: real-time video feeds
* Glue Data Catalog & Crawlers: Metadata repositories for schemas and datasets in your account
* Glue ETL: ETL Jobs as Spark programs, run on a serverless Spark Cluster
* DynamoDB: NoSQL store
* Redshift: Data Warehousing for OLAP, SQL language
* Redshift Spectrum: Redshift on data in S3 (without the need to load it first in Redshift)
* RDS / Aurora: Relational Data Store for OLTP, SQL language
* ElasticSearch: index for your data, search capability, clickstream analytics
* ElastiCache: data cache technology
* Data Pipelines: Orchestration of ETL jobs between RDS, DynamoDB, S3. Runs on EC2 instances
* Batch: batch jobs run as Docker containers - not just for data, manages EC2 instances for you
* DMS: Database Migration Service, 1-to-1 CDC replication, no ETL
* Step Functions: Orchestration of workflows, audit, retry mechanisms

Briefly mentioned, covered by Frank Kane:

* EMR: Managed Hadoop Clusters
* Quicksight: Visualization Tool
* Rekognition: ML Service
* SageMaker: ML Service
* DeepLens: camera by Amazon
* Athena: Serverless Query of your data