DIGITAL INNOVATION ONE

Surfando na Diversidade de Dados com AWS

Cassiano Peres
Analista e desenvolvedor de sistemas

DIGITAL INNOVATION ONE

Mais sobre mim

- Graduado em TADS – UTFPR-MD (2015)
- CTO – Arabyka e Brexbit
- Motivado pela liberdade e descentralização

linkedin.com/in/peres-cassiano/

github.com/cassianobrexbit

DIGITAL INNOVATION ONE

Desafio da Live

- Conceitos de ingestão, transformação e análise análise de dados;
- Ferramentas da AWS: SnowBall, Kinesis, MapReduce, Glue e Redshift
- Implementação uma ferramenta de análise utilizando AWS Kinesis, S3 e Glue DataBrew

DIGITAL INNOVATION ONE

Percurso

Etapa 1 Relembrando Big Data

Etapa 2 Ferramentas de Big Data na AWS

Etapa 3 Prática

DIGITAL INNOVATION ONE

Requisitos

- ✓ Conhecimento básico de Python, SSH e linux
- ✓ Conta na AWS
- ✓ Curiosidade e criatividade

DIGITAL INNOVATION ONE

Relembrando

Big Data trata dos desafios de gerenciamento de dados que não podem ser resolvidos com bancos de dados tradicionais devido aos crescentes:

- **Volume**: varia de terabytes a petabytes de dados
- **Variedade**: dados de ampla variedade de origens e formatos
- **Velocidade**: dados coletados, armazenados, processados e analisados em curtos períodos de tempo

DIGITAL INNOVATION ONE

Relembrando

A falha em tratar corretamente dos desafios de big data causam:

- Escalada de custos
- Redução de produtividade e competitividade.

Uma estratégia sólida de big data pode ajudar as organizações a:

- Reduzir custos
- Ganhar eficiência

DIGITAL INNOVATION ONE

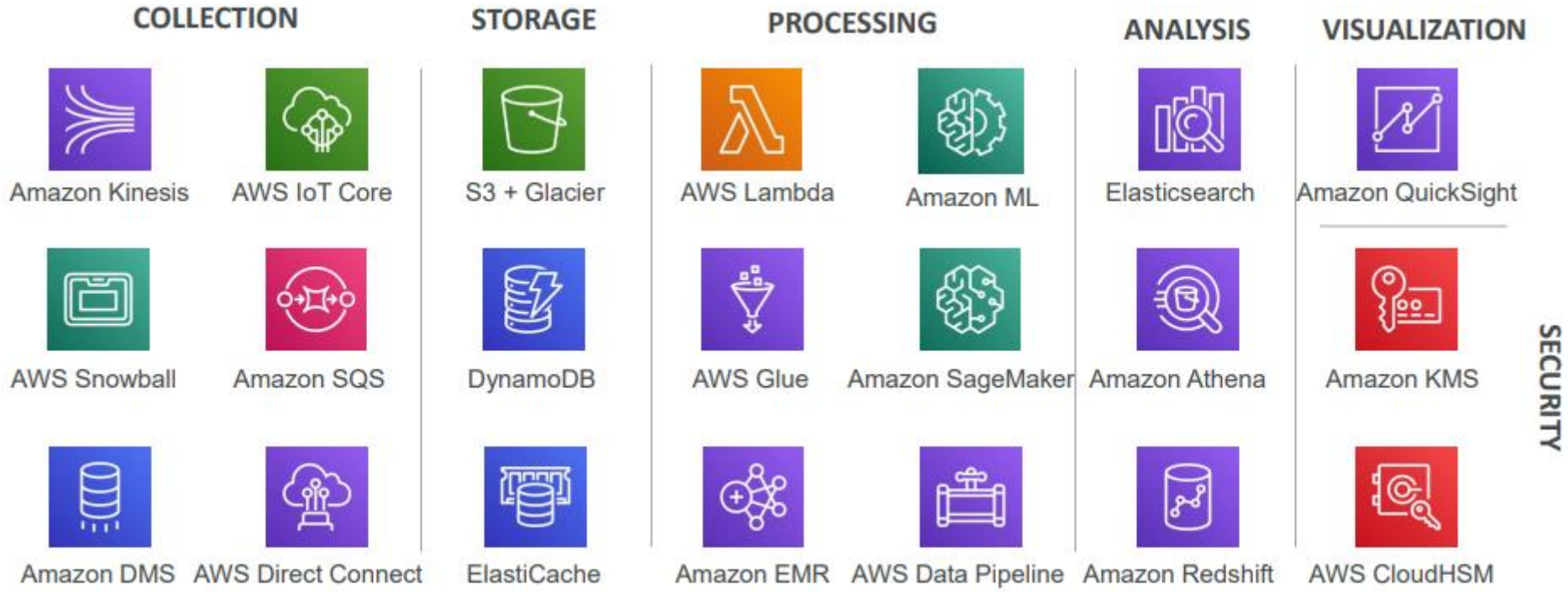
Relembrando

Na maioria dos casos, o processamento de big data envolve um fluxo de dados comum, da coleta de dados brutos ao consumo de informações práticas.

- **Coletar** dados brutos
- **Armazenar** de forma segura e escalável
- **Processar e analisar** transformando os dados
- **Consumir e visualizar** de forma amigável e agregando valor

DIGITAL INNOVATION ONE

Big Data na AWS



DIGITAL INNOVATION ONE

AWS Snow Family

- Dispositivos portáteis altamente seguros para coletar e processar dados e migração para dentro e para fora da AWS
- Dispositivos offline para realizar migrações de dados.

- Migração

- Edge computing

DIGITAL INNOVATION ONE

Aplicações

- Conectividade limitada
- Banda limitada
- Alto custo da rede
- Banda compartilhada reduz taxa de transferência
- Conexão instável

	Time to Transfer		
	100 Mbps	**1Gbps**	**10Gbps**
10 TB	12 days	30 hours	3 hours
100 TB	124 days	12 days	30 hours
1 PB	3 years	124 days	12 days

DIGITAL INNOVATION ONE

AWS Snow Family

Snowcone

Snowball Edge

Snowmobile

	Snowcone	Snowball Edge Storage Optimized	Snowmobile
Storage Capacity	8 TB usable	80 TB usable	< 100 PB
Migration Size	Up to 24 TB, online and offline	Up to petabytes, offline	Up to exabytes, offline
DataSync agent	Pre-installed		
Storage Clustering		Up to 15 nodes	

DIGITAL INNOVATION ONE

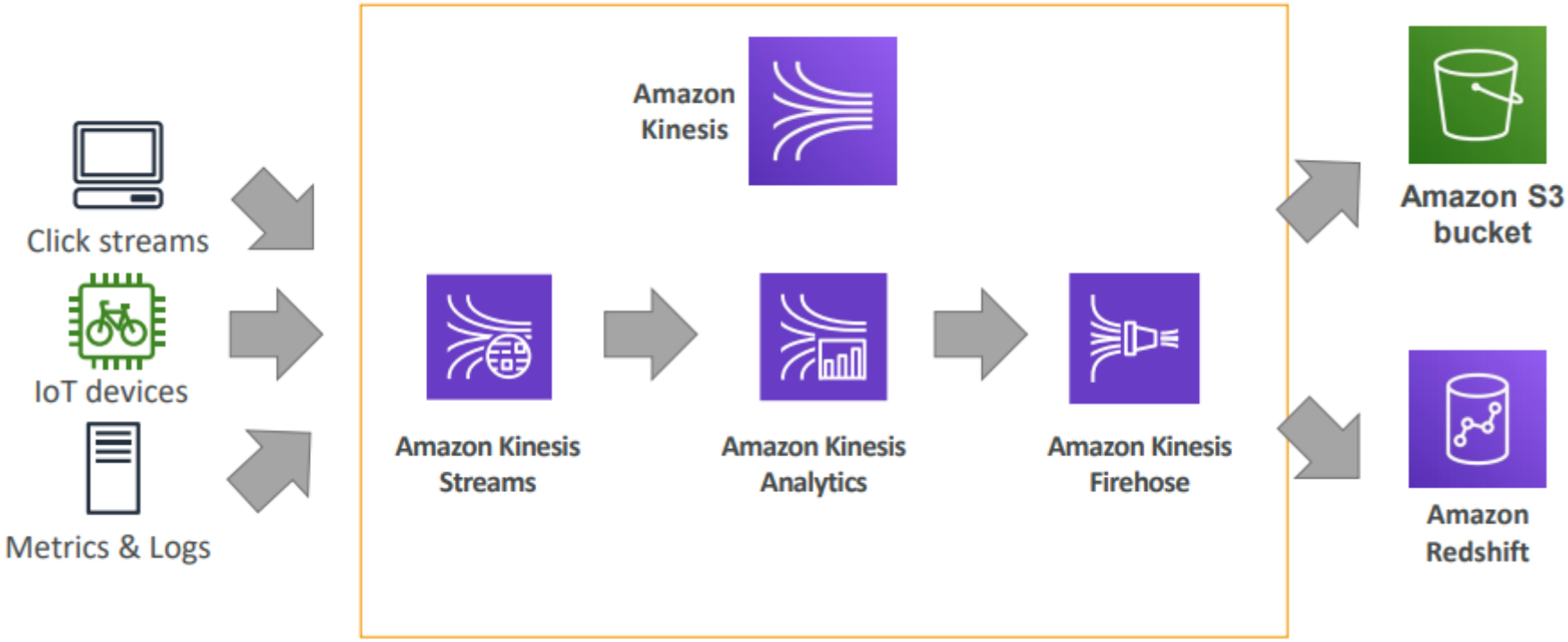
AWS Kinesis

Alternativa gerenciada ao Apache Kafka

- Ótimo para logs de aplicativos, métricas, IoT, clickstreams, big data "em tempo real“ e frameworks de processamento de streaming
- Dados replicados automaticamente de forma síncrona para 3 AZ
- **Kinesis Streams**: ingestão de streaming de baixa latência em escala (vídeo e data streams)
- **Kinesis Analytics**: análises em tempo real em streams usandoSQL
- **Kinesis Firehose**: fluxos de carga em S3, Redshift, ElasticSearch e splunk

DIGITAL INNOVATION ONE

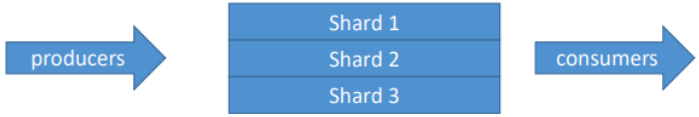
AWS Kinesis



DIGITAL INNOVATION ONE

Kinesis Streams

- Os fluxos são divididos em fragmentos (shards)/partições ordenados
- A capacidade total de um stream é a soma da capacidade dos shards
- Taxa de ingestão: 1MB ou 1000 mensagens/segundo
- Taxa de leitura: 2MB/s
- A retenção de dados é de 24 horas por padrão, pode ir até 7 dias
- Capacidade de reprocessar/reproduzir dados
- Vários aplicativos podem consumir o mesmo fluxo
- **Partition Key**: agrupar dados por shard em um stream
- **Sequence Number**: número único por partition key dentro de um shard



DIGITAL INNOVATION ONE

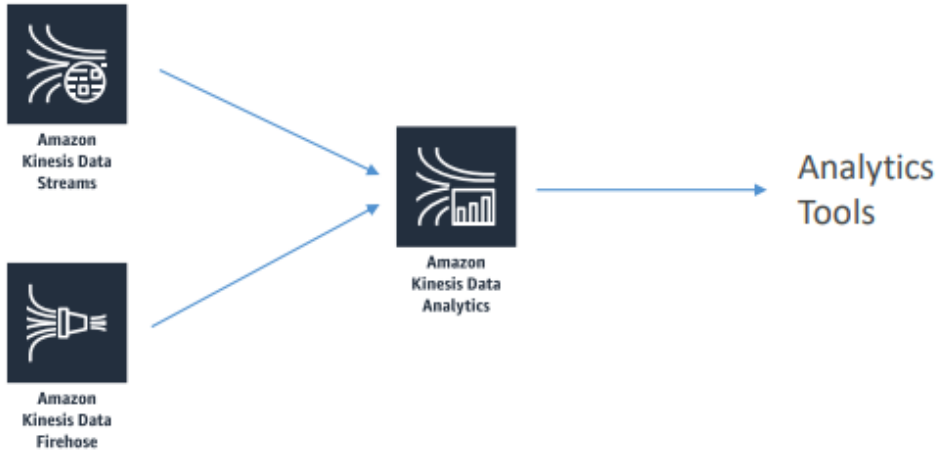
Kinesis Streams

- **Kinesis Producers**: aplicativo que insere registros de dados de usuário em um stream de dados do Kinesis – Ingestão.
- **Kinesis Consumers** : aplicação para ler e processar registros de dados de streamings de dados do Kinesis.
- **Kinesis Agent**: monitora continuamente um conjunto de arquivos e envia novos dados ao stream. Baseado em Java.

Kinesis Analytics

O Amazon Kinesis Data Analytics é a maneira mais fácil de transformar e analisar dados de streaming em tempo real com o Apache Flink (estrutura e um mecanismo de código aberto para o processamento de streams de dados).

- Streaming ETL
- Geração contínua de métricas
- Análise responsiva



DIGITAL INNOVATION ONE

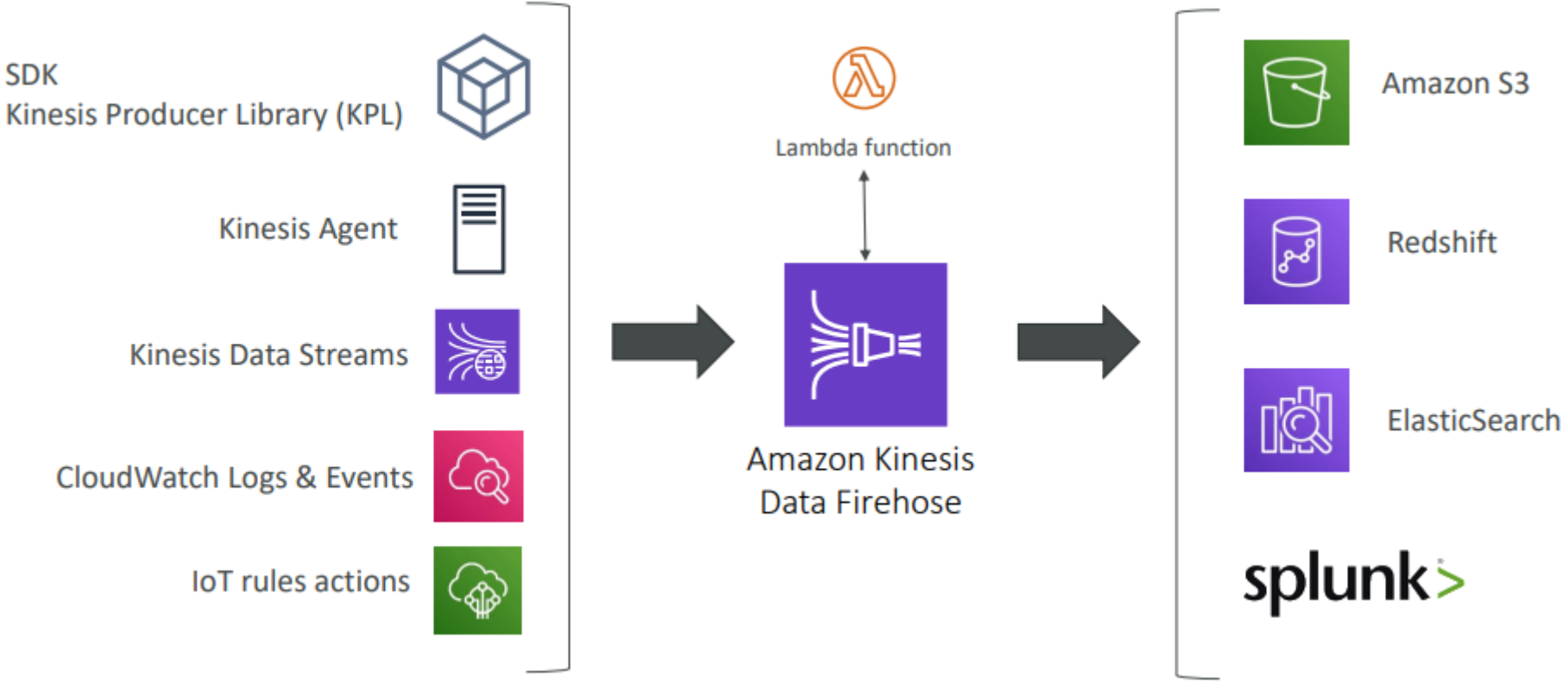
Kinesis Data Firehose

Serviço totalmente gerenciado, sem administração

- Quase em tempo real (latência de 60 segundos)
- Carregar dados no Redshift / Amazon S3 / ElasticSearch / Splunk
- Escala automática
- Suporta diversos formatos de dados

DIGITAL INNOVATION ONE

Kinesis Data Firehose



AWS Elastic MapReduce

Serviço totalmente gerenciado

- Estrutura gerenciada do Hadoop em instâncias EC2
- Inclui Spark, HBase, Presto, Flink,Colmeia e mais
- Notebooks EMR
- Vários pontos de integração com AWS
- HDFS

AWS Elastic MapReduce

Cluster EMR

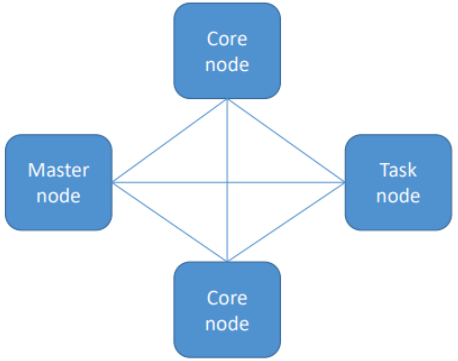
Master node: gerencia o cluster

- Rastreia o status das tarefas, monitora o cluster
- Única instância EC2
- Também conhecido como “nó líder”

Core node: hospeda dados HDFS e executa tarefas

Task Node: executa tarefas, não hospeda dados

- Opcional
- Sem risco de perda de dados ao remover
- Bom uso de instâncias pontuais



DIGITAL INNOVATION ONE

AWS Glue

AWS Glue é um serviço de integração de dados **sem servidor** que facilita descobrir, preparar e combinar dados para análise, machine learning e desenvolvimento da aplicação.

- **AWS Glue DataBrew**: ferramenta visual para enriquecer, limpar e normalizar os dados sem escrever código.
- **AWS Glue Elastic Views**: permite utilizar SQL (Structured Query Language) para combinar e replicar os dados em diferentes armazenamentos de dados.

DIGITAL INNOVATION ONE

AWS Glue

Glue Crawler verifica dados em S3 e cria esquemas

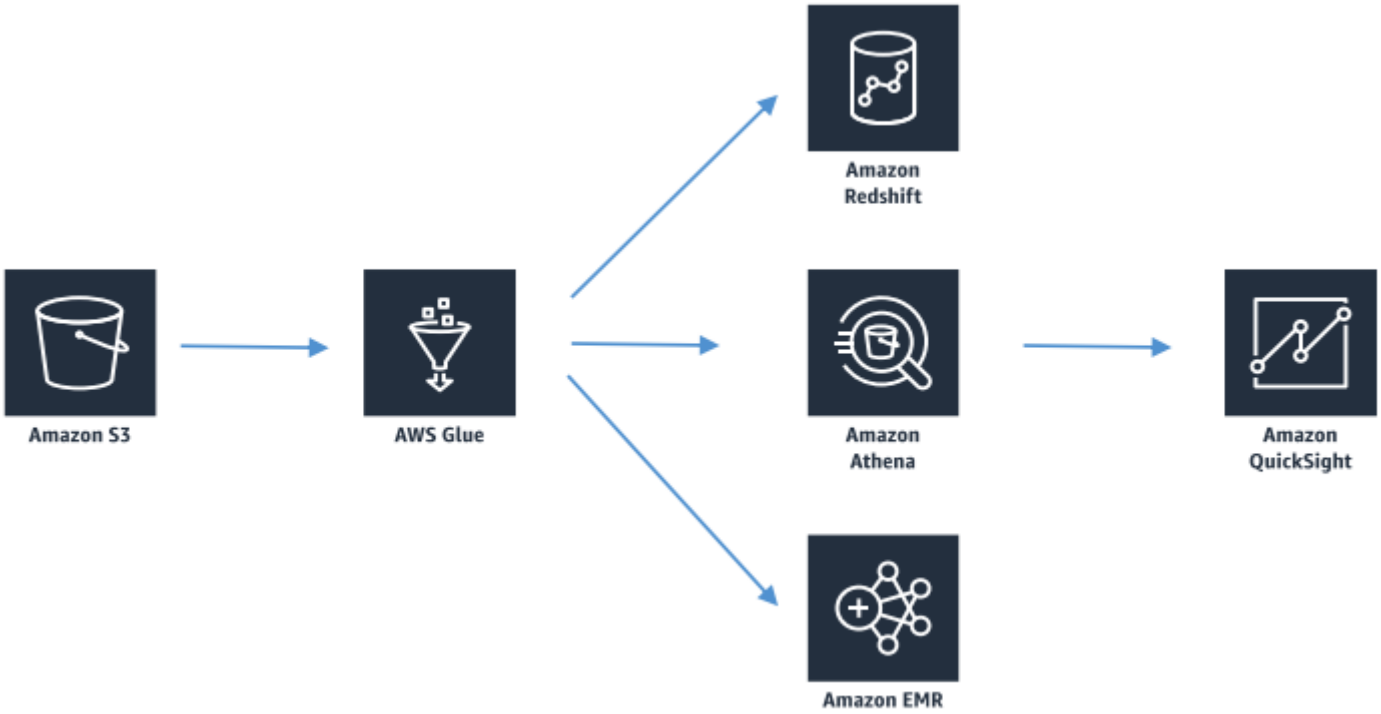
- Pode ser executado periodicamente
- Preenche o Catálogo de Dados do Glue
- Armazena apenas a definição da tabela
- Os dados originais permanecem no S3

Dados não estruturados são tratados como estruturados

- Redshift Spectrum
- Athena
- EMR
- Quicksight

DIGITAL INNOVATION ONE

AWS Glue



DIGITAL INNOVATION ONE

AWS Redshift

Serviço de armazenamento de dados gerenciado em escala de petabytes.

- desempenho 10 vezes melhor do que outros DW
- Projetado para OLAP, não OLTP
- Econômico
- Interfaces SQL, ODBC, JDBC
- Escala sob demanda
- Replicação e backups integrados
- Monitoramento via CloudWatch / CloudTrail

Amazon Redshift

DIGITAL INNOVATION ONE

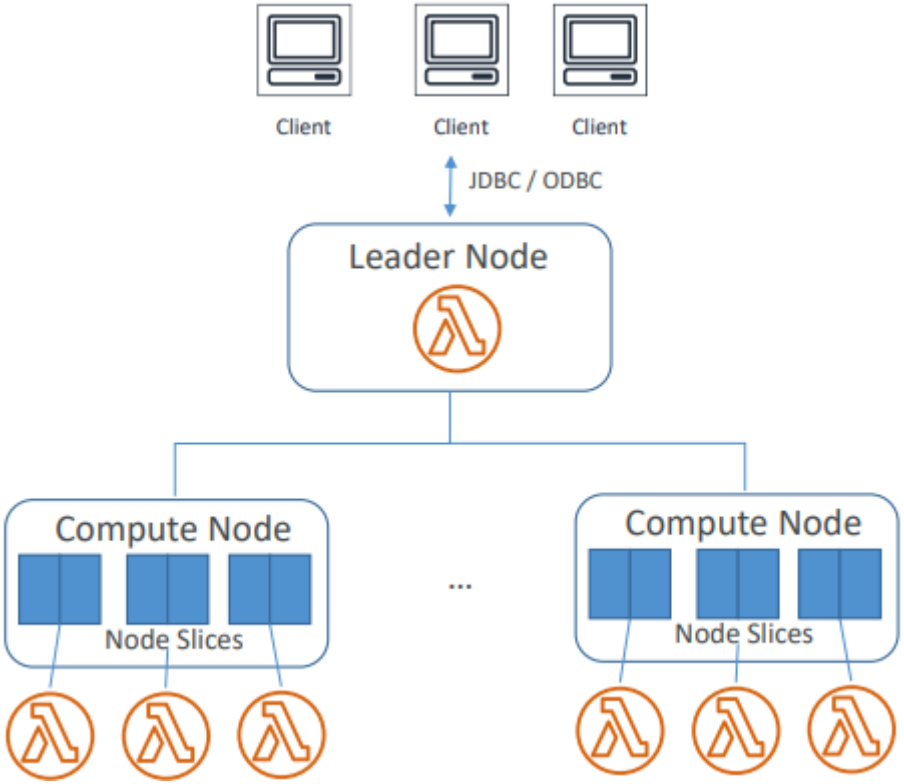
AWS Redshift

Usos

- Acelerar as cargas de trabalho de análise
- Data warehouse e data lake unificados
- Modernização do data warehouse
- Analisar dados de vendas globais
- Armazenar dados históricos do mercado de ações
- Analisar impressões de anúncios e cliques
- Dados de jogos agregados
- Analisar tendências sociais

DIGITAL INNOVATION ONE

AWS Redshift



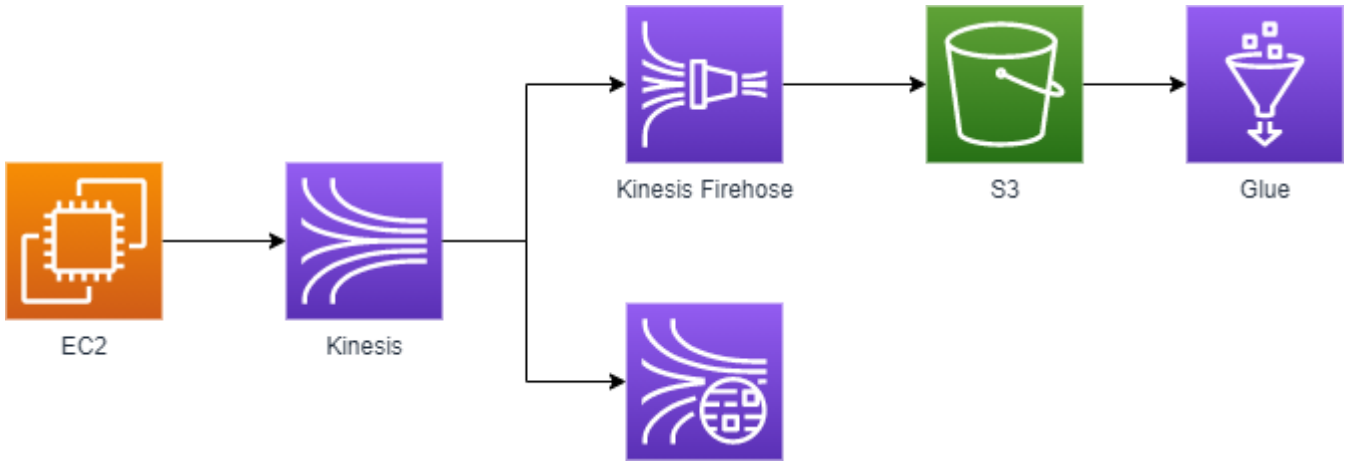
DIGITAL INNOVATION ONE

Mãos à obra

- Criar uma Stream Delivery com o AWS Kinesis Firehose
- Configurar instância no AWS EC2
- Gerar logs de processamento de dados com Python
- Armazenar logs no AWS S3
- Manipular dados no AWS Glue Data Brew

DIGITAL INNOVATION ONE

Arquitetura da prática



DIGITAL
INNOVATION
ONE

Dúvidas?

DIGITAL INNOVATION ONE

Referencial

GitHub: https://github.com/cassianobrexbit/dio-live-aws-bigdata-2.git
Snow: https://aws.amazon.com/pt/snow/
Kinesis: https://aws.amazon.com/pt/kinesis/
Glue: https://aws.amazon.com/pt/glue/
EMR: https://aws.amazon.com/pt/emr/