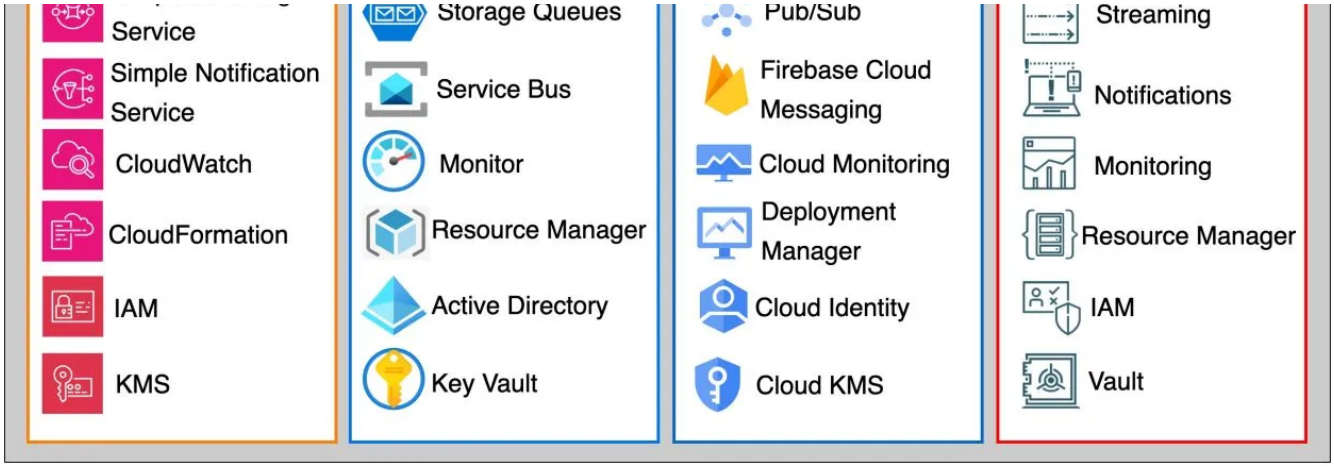
LESA6

Os grandes provedores de serviços de computação em nuvem (hyper scalers) disponibilizam soluções que vão desde armazenamento e processamento até cumprimento a normas de regulamentação.

Como as soluções atendidas são todas relacionadas a computação, é possível fazer paralelos entre as soluções apresentadas por cada provedor, conforme apontado por Govardhana Miriyala Kannaiah em publicação do blog bytebytego: A nice cheat sheet of different cloud services (2023 edition)



Exemplos de serviços utilizados com frequencia por seus diferenciais

Considerando a capacidade de distribuição de computação, os seguinte serviços são utilizados com bastante frequência considerando as vantagens e desvantagens de suas respectivas descriçoes:

Virtual machine

Com o desenvolvimento de tecnologias de computação em grid abre-se a possibilidade de fracionamento do poder de computação em diferentes configurações. Assim, é possível utilizar instâncias de máquina virtual que rodam sobre a infraestrutura do hyper scaler de acordo com a necessidade de negócio.

Considerando a gama de serviços disponíveis, é uma solução menos flexível e sujeita a custos mais constantes.

Azure Kubernetes Service (AKS)

Um dos grandes potenciais da nuvem é o emprego de containeres efêmeros (que podem ser criados e destruídos de acordo), por permitir que clusteres de computadores possam ser orquestrados como solução elástica através de tecnologias como kubernetes.

Entretanto, gerenciar os containeres de controle é uma tarefa que exige contratação de profissionais qualificados e escassos no mercado, bem como ter necessidade computacional que justifique a aquisição de tal maquinário para ter disponível uma nuvem privada.

Dessa forma, com a utilização de serviços gerenciados de kubernetes, é possível ter inúmeras aplicações sendo executadas, implantadas e removidas da operação sem necessidade de mudanças significativas na infra estrutura.5

Azure functions

Abstraindo ainda além do kubernetes, o conceito de servico gerenciado pode ser implementado através da tecnologia Azure Function. Essa tecnologia permite que se defina o poder de processamento, volume de memória e o código em linguagem de programação que a ser executado.

Essa é uma das tecnologias denominadas como serverless, por passar a responsabilidade do gerenciamento de hardware, sistema operacional ou de redes. Assim, o negócio ser preocupará apenas com a lógica a ser implementada.

Blob Storage

É uma forma de armazenamento em nuvem que permite que diferentes tipos de objetos sejam armazenados. Diferentemente de um banco de dados relacional, os dados são armazenados sequencialmente sem organização de tupla. Esse tipo de organização permite que os objetos sejam recuperados mais rapidamente.

Também são distribuídos entre as zonas de disponibilidade, garantindo a aplicação de redundância para a operação dos serviços, respeitando questões legais e de governança.

Considerando a velocidade de acesso de memória e escalabilidade que os serviços gerenciados podem alcançar, é um forte candidato para leitura e armazenamento de processamentos paralelos.

Cosmos DB

É um banco de dados NoSql. Diferentemente do Blob Storage, permite que dados sejam acessados sem a necessidade de recuperar o objeto inteiro primeiro. Também é um banco de dados escalável e com orientação à performance para aplicações. A performance do banco é relacionado à indexação dos atributos dos objetos. O armazenamento dos atributos podem ser configurados para melhoria de performance. Algumas configurações como TTL (Time to live), são boas práticas para melhor controle de custos.

Event hub

É uma tecnologia de fila, desenvolvida com foco em processamento de eventos em tempo real de grandes volumes de dados e diversidade de origens.

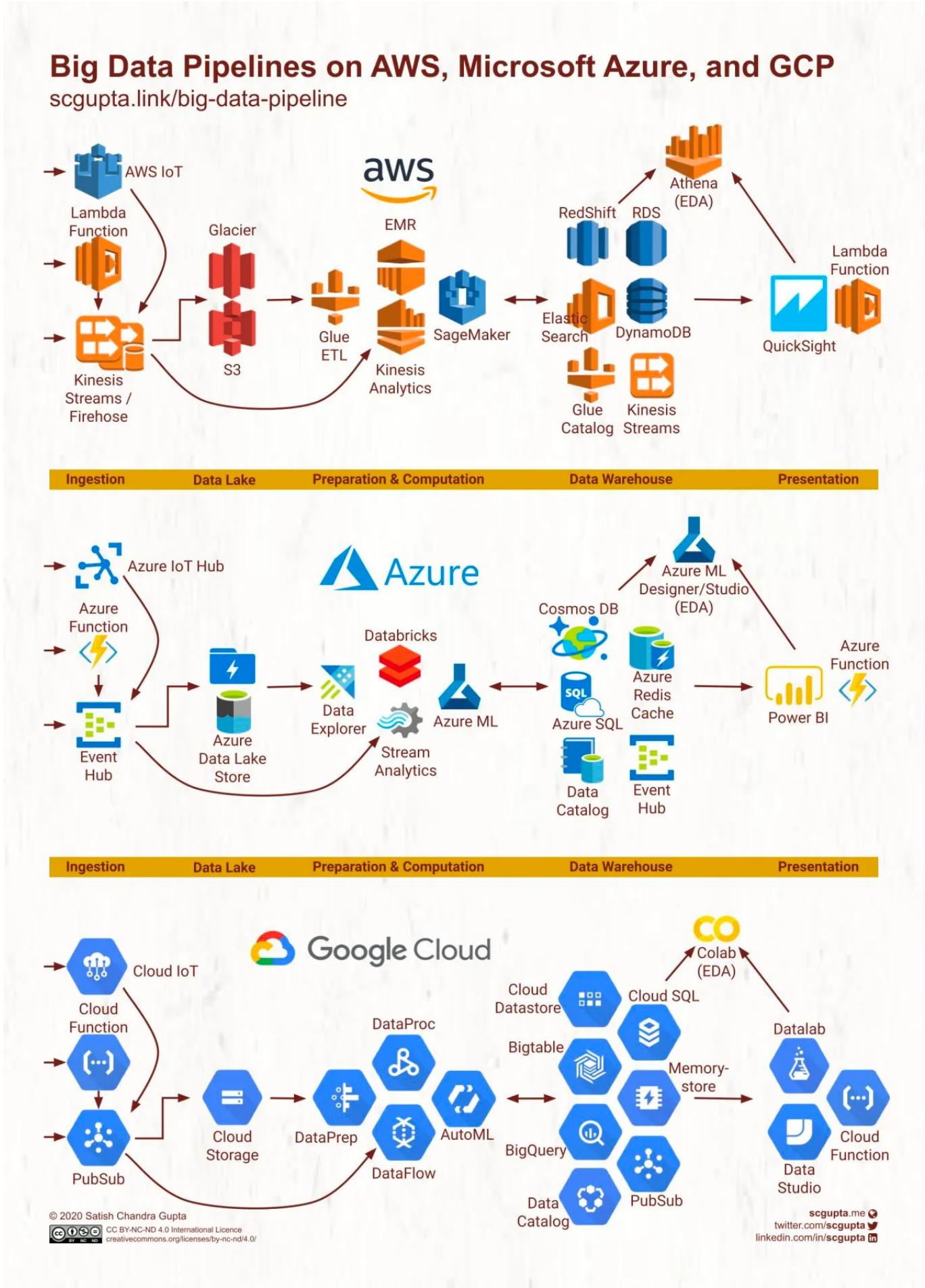
A tecnologia permite que aplicações stateless possam ser expandidas horizontalmente sem disputa de recursos ou duplicidade de itens.

Exemplo prático de aplicação dos serviços

Tratando sobre exemplos de escalabilidade, propostas de soluções para big data são propícios por apresentarem as seguintes características:

- Grandes volumes de dados de diferentes origens e estruturas
- O processamento precisa ocorrer em determinado período ou continuamente
- O tamanho dos nós de processamento podem exigir arquiteturas que possam escalar horizontalmente ou verticalmente, de acordo com o tipo de situação a ser solucionada

Novamente, trazemos um exemplo comparado do blog bytebytego em que é apresentada a mesma sugestão de solução implementada em diferentes hyper scalers: Which cloud provider should be used when building a big data solution?



Dessa forma, abordaremos a possibilidade de implementação de solução no hyper scaler da Microsoft, Azure cloud.

Ingestão

Na etapa de ingestão, é sugeriada a possibilidade de utilização dos serviços Azure Iot Hub e Azure Function se comunicando com o serviço Event Hub.

A tecnologia Event hub é capaz de receber eventos de diversas origens de dados e disponibilizar em tempo real em ordem de chegada para que seja consumido no esquem Pub/Sub

Data Lake

Soluções de data lake são projetada para suportar persistência de grandes volumes de dados exigem grande largura de banda, capacidade de armazenamento e método de armazenamento capaz de suporta demanda de diferentes pontos em alta disponibilidade.

Serviços de datalake têm o objetivo de disponibilizar dados para que possam ser processados por outros processos, sendo normalmente focados em armazenar dados brutos.

O serviço gerenciado Azure Data Lake Store permite que grandes volumes de dados possam ser armazenados e consumidos em série. Não é adequado para acessar dados de forma transacional, por ter design voltado para performance de consumo de grandes volumes de dados.

Uma vantagem é que não é requerida licença para utilização da ferramenta, sendo pago apenas o que for consumido.

Preparation and Computation

São listadas diversas soluções que podem ser empregadas em combinação ou isoladamente:

- Azure Databricks: É uma solução de plataforma de dados que implementa o serviço da empresa de mesmo nome com foco em capacidade e machine learning. Roda sobre a solução Apache Spark, sendo compatível com Hadoop, R e SQL, por exemplo.
- Data explorer: É um serviço com capacidades de monitoramento em tempo real, utilizado em situações de negócios ou IoT, por exemplo. Permite explorar os dados em tempo real no dashboard integrado.
- Azure Stream Analytics: É uma ferramenta focada em facilidade de uso, no code e análise em tempo real para cargas de trabalho críticas.
- Azure Machine Learning (Azure ML): É uma ferramenta que permite que projetos de machine learning possam ser colocados em produção mais rápido através de práticas de MLOps. Para que os treinamentos sejam implementados em velocidade satisfatória, é importante que tanto o modelo sendo treinado tenha os recursos de processamento disponíveis quanto memória e banda.

Data Warehouse

Tanto quanto capacidade de processamento de dados, a capacidade de persistir dados sem perda de velocidade ou qualidade são cruciais para que não se perca as informações de etapas anteriores.

Soluções de persistência de dados armazenam tanto dados em bancos relacionais quanto não relacionais cumprindo a tarefa de confiabilidade. através de diferentes estratégias:

- Cosmos DB: É um banco de dados NoSql que permite que os dados sejam armazenados de forma a guardar apenas dados de range, ou atributos, ganhando velocidade na recuperação de dados prontos para processamento.
- Azure Sql Database: É um banco de dados multi modelo, é projetado para ter integração com os demais serviços da hyper scaler, suportanto acessos de diferentes serviços.
- Azure Redis: É um banco de dados em memória, originalmente empregado para cache. Implementações modernas em formato de serviço oferencem redundâncias para o banco, resolvendo questões de volatilidade de bancos de dados em memória, sendo um forte candidato para resolver situações em que baixa latência e capacidade de escrita são importantes.
- Event Hub: Assim como na ingestão, permite que os dados sejam armazenados em uma stream de eventos até que uma nova etapa possa consumir os dados adicionados nesse momento.

Presentation

As soluções de apresentação permitem que os dados já processados sejam exibidos.

Para a solução proposta, não são necessárias grandes escalabilidades, por se tratar de uma perpectiva que consome dados já preparados para projeção.