

BIG DATA: PROPOSTA DE UM DATA WAREHOUSE PARA APOIO À TOMADA DE DECISÃO EM UMA COMPANHIA DE ABASTECIMENTO DE ÁGUA E SANEAMENTO BÁSICO

Carlos Henrique Batista Martins¹, Fabio Lopes Junior², Luis Marcelo Bortolotti³

¹ Discente do Curso Superior de Tecnologia em Análise e Desenvolvimento de Sistemas/ carlos.martins18@fatec.sp.gov.br

² Discente do Curso Superior de Tecnologia em Análise e Desenvolvimento de Sistemas / fabio.lopes21@fatec.sp.gov.br

³ Docente do Curso Superior de Tecnologia em Análise e Desenvolvimento de Sistemas / luis.bortolotti@fatec.sp.gov.br

RESUMO

Este trabalho foi feito com base em pesquisas e tem o objetivo de analisar ferramentas para trabalhar com grandes quantidades de dados, o conceito que chamamos de Big Data. Uma das ferramentas que destacamos é o modelo de banco de dados Data Warehouse, que através da sua abordagem chamada de esquema estrela oferece simplicidade, capacidade de análise multidimensional e facilidade de manutenção dos dados. O objetivo desse projeto é mostrar que com o uso de um data warehouse é possível realizar análise mais profundas e mais precisa, mesmo quando se trata de dados em grande quantidade e diversidade, assim trazendo mais eficiência e mais desempenho nos mais diversos setores do mercado.

Palavras-chave: big data; ferramentas; data warehouse; esquema estrela.

1 INTRODUÇÃO

Com a evolução constante da tecnologia o volume de dados e da informação vem crescendo de forma acelerada, o que gera para as instituições comerciais o grande desafio de gerenciar uma quantidade de dados cada vez maior, e nem sempre essas instituições têm condições de tratar esses dados de forma adequada a fim de usá-los de forma eficiente.

É a capacidade de trabalhar com essa enorme quantidade de dados que chamamos de Big Data, e é assim que Thomas H. Davenport, professor de Tecnologia e Gestão da Informação na Babson College define o termo: “Big Data é o processo de encontrar insights significativos em grandes quantidades de dados. Isso envolve o uso de técnicas analíticas sofisticadas para extrair informações úteis dos dados.”

Empresas como Google e Meta podem gerar quantidade de dados na casa dos exabytes (EB), o que gera dificuldade para processamento desses dados, um exemplo disso são os dados gerados por dispositivos IoT (Internet das Coisas) que não são coletados ou processados de forma eficiente, pois não há uma infraestrutura adequada para coletar e analisar esses dados. Além disso, muitos dados gerados por redes sociais, transações financeiras e outras fontes podem ser de baixa qualidade, incompletos ou irrelevantes, o que pode dificultar seu uso efetivo.

E é aí que pode entrar um Data Warehouse, que se trata de um sistema de gerenciamento capaz de coletar, armazenar e gerenciar grandes quantidades de dados com o objetivo de análise para negócios. É projetado para trabalhar com dados de vários tipos e de várias fontes diferentes com o intuito de potencializar a tomada de decisões das empresas com base nesses dados coletados.

2 REFERENCIAL TEÓRICO

2.1 Big Data

Segundo Davenport (2014) e Marr (2015), o big data se refere à enorme quantidade de dados estruturados e não estruturados - que inundam uma empresa diariamente. Mas não é a quantidade de dados que é importante, são o que as organizações fazem com os dados que conta. Big data pode ser analisado para insights que levam a melhores decisões e movimentos estratégicos. Um insight é uma percepção clara sobre determinada situação.

Tomando como base essas definições, o big data é referente a enorme quantidade de dados gerados e coletados todos os dias, no qual são provenientes de fontes como as redes sociais, sensores, transações financeiras, registros médicos, entre outras. Gerenciar todos esses dados em imensa quantidade é na maioria das vezes um grande desafio para as empresas e organizações de todo o mundo e de todos os setores. E por conta desse desafio, o Big Data tem sido um tema cada vez mais difundido por vários setores, dentre eles saúde, finanças, varejo, manufatura, entre outros. Uma de suas principais vantagens é a capacidade de obter insights mais profundos e significativos sobre padrões, tendências e comportamentos, que tendem a ajudar as empresas a tomar certas decisões mais informadas e estratégicas. Com essas ferramentas de análise avançadas, como machine learning e inteligência

artificial, é possível identificarmos correlações e padrões que não seriam visíveis com outras técnicas de análises que são tradicionais.

Ainda segundo Davenport (2014), o conceito de análise de dados, que é a base para o big data, é utilizado pelo menos desde 1954, quando a empresa de transportes, *United Parcel Service* (UPS), criou um grupo de *analytics*. A tabela (1) apresenta um histórico de termos já utilizados para definir esse conceito.

Terminologia para uso e análise de dados		
Termo	Período	Significado específico
Suporte a decisão	1970 - 1985	Uso da análise de dados para dar suporte a tomada de decisões.
Suporte a executivos	1980 - 1990	Foco na análise de dados para dar suporte ao processo decisório dos altos executivos.
Processamento analítico on-line (OLAP)	1990 - 2000	Software para análise de tabela de dados multidimensionais.
Business Intelligence	1989 - 2005	Ferramentas para dar suporte a decisões orientadas por dados, com ênfase em relatórios.
Analytics	2005 - 2010	Foco em análises estatísticas e matemáticas para a tomada de decisões.
Big Data	2010 - Atualmente	Foco em um grande volume de dados não estruturados e rápido movimento.

Tabela 1: Histórico de uso de análise de dados. Adaptado de Davenport (2014)

Dentre os principais princípios do big data destacam-se os que são chamados de “Os V’s do Big Data” que são eles (UFRJ, 2015):

Volume: É mais observado em fatos cotidianos: durante o dia, com o volume de troca de e-mails, transações bancárias, interações nas redes sociais, registro de chamadas e tráfego de dados nas linhas telefônicas. Esses fatos nos passam uma breve noção de dados que existem atualmente. É estimado que o volume total de dados que temos na internet por ano é de 250 Exabytes(10^{18} bytes).

Velocidade: Segundo Taurion (2014) a velocidade é tão importante que num dado momento deverá existir uma ferramenta capaz de analisar os dados em tempo real. Hoje em dia, essa análise só acontece após os dados serem armazenados, porém com o tempo gasto no armazenamento, já desclassifica esse tipo de análise como uma análise 100% em tempo real. Informação é poder sendo a velocidade a qual você tem essa informação é uma vantagem competitiva das empresas. A velocidade pode acabar limitando a operação de muitos negócios, por exemplo quando usamos um cartão de crédito, se não é aprovado a compra em segundos, já se pensa em utilizar outro método de pagamento, aí a operadora perde uma oportunidade pela falha na velocidade de sua transmissão e análise de dados.

Variedade: Com o volume de dados que temos, adquirimos também uma variedade dos mesmos. Pode-se observar a variedade de dados nos e-mails, redes sociais, fotografias, áudios, telefones e cartões de crédito. Empresas que conseguem captar essa variedade, seja de critérios ou por fontes, conseguem agregar mais valor ao negócio .O Big Data escalona a variedade de informações na seguinte forma:

- Dados estruturados: Armazenados em bancos de dados
- Dados semi-estruturados: Acompanham padrões heterogêneos, mais difíceis de serem identificados por seguir vários padrões.
- Dados não estruturados: São uma mistura de dados com fontes diversificadas, como áudios, imagens e documentos online.

Dentro das 3 categorias, estima-se que até 90% de todos os dados no mundo estão na forma de dados não estruturados.

Veracidade: Um a cada 3 líderes não confiam nos dados que recebem. Para ter bons resultados com o processo do Big Data é necessário obter dados verídicos. A velocidade já explicada é bem alinhada junto ao conceito de veracidade por causa da necessidade constante das análises em tempo real, de dados que condiz com a realidade do momento, pois os dados passados não podem mais ser considerados dados verídicos para o momento no qual é analisado. A relevância dos dados

coletados é tão importante quanto o primeiro conceito, a verificação dos dados coletados para adequação e relevância ao propósito da análise acaba sendo um ponto chave para obter dados que agreguem valor ao processo.

Valor: Quanto maior a riqueza de dados, mais importante é saber realizar as perguntas corretas no começo de todo processo de análise. É necessário focar para a orientação do negócio, o valor que a coleta e análise dos dados trará para o negócio. Não é viável realizar todo o processo do Big Data se não tem questionamentos que ajudem o negócio no modo realístico.

2.1.1 Tecnologias para Big Data

Existem muitas tecnologias que podem ser aplicadas ao conceito de big data dentre elas se destacam:

Hadoop é um *framework* de código aberto desenvolvido para armazenamento e processamento distribuído de grandes conjuntos de dados em cluster de computadores. Foi criado para lidar com o desafio de lidar com a crescente quantidade de dados gerados pelas organizações e empresas. É baseado no modelo de programação *MapReduce*, que divide tarefas em blocos pequenos de trabalho e os distribui em diferentes nós de um cluster de computadores, no qual permite que grandes conjuntos de dados sejam processados em paralelo, acelerando o tempo do processamento. Além do modelo *MapReduce*, o *Hadoop* inclui o *Hadoop Distributed File System* (HDFS) que é um sistema de arquivos distribuído que fornece armazenamento de arquivos de grande escala para os dados processados pelo Hadoop. Ele é utilizado por muitas empresas e organizações para lidar com grandes quantidades de dados em tempo real, como empresas de tecnologia, finanças, saúde, governo e outras indústrias. (SAS Institute, 2020)

Cloud Computing é um modelo de fornecimento de serviços de computação sob demanda, que permite o acesso a recursos de computação compartilhados, como servidores, armazenamento, aplicativos e serviços, por meio da internet. Os serviços de computação em nuvem são oferecidos em diferentes modelos, incluindo *Software as a Service* (SaaS) *Platform as a Service* (PaaS) e *Infrastructure as a Service* (IaaS). O SaaS oferece aplicativos prontos para uso, como softwares de e-mail e de gerenciamento de projetos. O PaaS oferece um ambiente de desenvolvimento e

execução para aplicativos personalizados, enquanto o IaaS oferece recursos de infraestrutura, como servidores virtuais e armazenamento em nuvem. A computação em nuvem oferece várias vantagens, como escalabilidade, flexibilidade, custos reduzidos e segurança. Essa flexibilidade permite que as empresas acessem seus recursos de computação de qualquer lugar do mundo, permitindo o trabalho remoto. O modelo de pagamento por uso significa que as empresas só pagam pelos recursos que usam, reduzindo os custos operacionais. A segurança é garantida por meio de recursos de segurança em camadas oferecidos pelos provedores de serviços em nuvem, incluindo criptografia, autenticação de usuários e backup de dados.(Amazon, 2021)

O Apache Kafka é uma plataforma de streaming de dados de código aberto desenvolvida pela Apache Software Foundation. É projetado para lidar com grandes volumes de dados em tempo real, permitindo que as empresas capturem, processem e transmitam fluxos de dados em tempo real entre diferentes sistemas e aplicativos. Ele é altamente escalável, tolerante a falhas e pode ser executado em clusters de servidores distribuídos, permitindo que as empresas gerenciem grandes volumes de dados em tempo real. O Kafka é frequentemente usado em cenários de Big Data, IoT (Internet das Coisas), análise de dados em tempo real e outras aplicações que exigem a transmissão de grandes volumes de dados em tempo real. Ele permite que as empresas capturem dados de diferentes fontes e transmitam esses dados para diferentes aplicativos e sistemas em tempo real. O Kafka também possui uma API rica e suporta várias linguagens de programação, permitindo que os desenvolvedores integrem facilmente o Kafka em seus aplicativos e sistemas existentes.(IBM, 2020)

2.2 Data Warehouse

Um data warehouse é um modelo de banco de dados que tem o objetivo de fornecer dados para consulta e tomada de decisões diferente de um banco de dados operacional, que fornecem dados para o funcionamento de determinado sistema. Em um data warehouse os dados são organizados por assuntos, por exemplo uma empresa de varejo, pode haver tabelas de fatos relacionadas às vendas, estoque ou logística. Cada uma dessas tabelas de fatos será projetada para armazenar as métricas específicas dessas áreas de negócio. Essa organização por assuntos permite que os usuários acessem facilmente os dados relevantes para suas análises específicas, focando em um determinado aspecto ou área de negócio. Isso ajuda a

simplificar as consultas e a geração de relatórios, tornando a análise mais eficiente e direcionada aos objetivos do usuário.(Oracle, 2023)

2.2.1 Data Warehouse x Data Mart

O data warehouse é um grande repositório centralizado que abrange toda a organização, enquanto o Data Mart é uma subdivisão especializada focada em uma área específica ou equipe. O DW fornece uma visão corporativa ampla, enquanto o Data Mart oferece uma visão mais direcionada e detalhada para atender às necessidades específicas dos usuários.(IBM, 2021)

2.2.2 Cubo de dados

Um cubo de dados é uma estrutura multidimensional que organiza e fornece acesso eficiente aos dados em um Data Warehouse, permitindo análises flexíveis e visuais. Um cubo de dados é composto por dimensões, medidas e hierarquias. As dimensões representam as características de negócio pelos quais os dados são analisados, como tempo, produto, cliente, região, entre outros. As medidas são os valores numéricos que são analisados, como vendas, lucro, quantidade, etc. A figura 1 ilustra o cubo de dados.

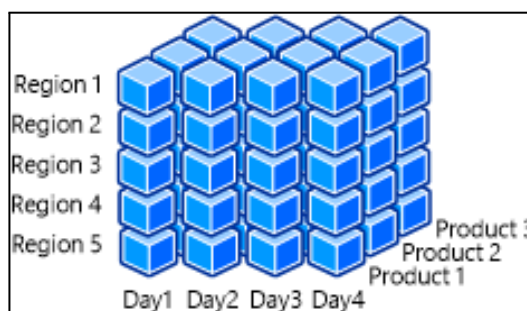


Figura 1: Cubo de dados (Microsoft, 2023)

2.3 Ambiente de data warehousing

Um ambiente de data warehousing é um sistema projetado para armazenar, gerenciar e analisar grandes volumes de dados empresariais. Ele consiste em um banco de dados centralizado, ferramentas de extração, transformação e carga (ETL), ferramentas de modelagem dimensional, ferramentas de consulta e relatórios, servidores de processamento, segurança e controle de acesso, bem como administração e monitoramento. Esse ambiente permite que as organizações

organizem e consolidem seus dados de várias fontes, tornando-os acessíveis para análise e tomada de decisões informadas.(Oracle, 2023)

2.3.1 Definindo ETL

Extrair, transformar e carregar (ETL) é o processo usado para coletar dados de várias fontes e reuni-los para dar suporte à análise e à tomada de decisões. As origens de dados podem ser muito diversas em formato, confiabilidade e volume, por tanto precisam ser processados para serem úteis quando reunidos, um processo que passa por três etapas distintas: A primeira é Extrair, durante a extração, o ETL identifica os dados e os copia de suas origens, de forma que possa transportar os dados para o armazenamento de dados de destino. A segunda etapa é Transformar, no processo de transformação, o ETL válido, autêntico e agrega os dados de formas que tornam os dados resultantes confiáveis e consultáveis. A última etapa é carregar, o ETL move os dados transformados para o armazenamento de dados de destino. Esta etapa pode implicar o carregamento inicial de todos os dados de origem ou pode ser o carregamento de alterações incrementais nos dados de origem.(SAS Institute, 2020)

2.3.2 Esquema estrela

O esquema em estrela é uma abordagem de modelagem madura amplamente adotada por data warehouses. Ele requer que os modeladores classifiquem suas tabelas de modelo como dimensão ou fato. Tabelas de dimensões descrevem as entidades de negócios. As entidades podem incluir produtos, pessoas, locais e conceitos, incluindo o próprio tempo. Uma tabela de dimensões contém uma ou mais colunas de chave, que atuam como um identificador exclusivo, e colunas descritivas. Tabelas de fatos armazenam observações ou eventos e podem ser ordens de vendas, saldos de ações, taxas de câmbio, temperaturas, etc. Uma tabela de fatos contém colunas chave de dimensões relacionadas a tabelas de dimensões. (Microsoft, 2023)

A figura 2 ilustra o conceito de modelagem do esquema estrela.

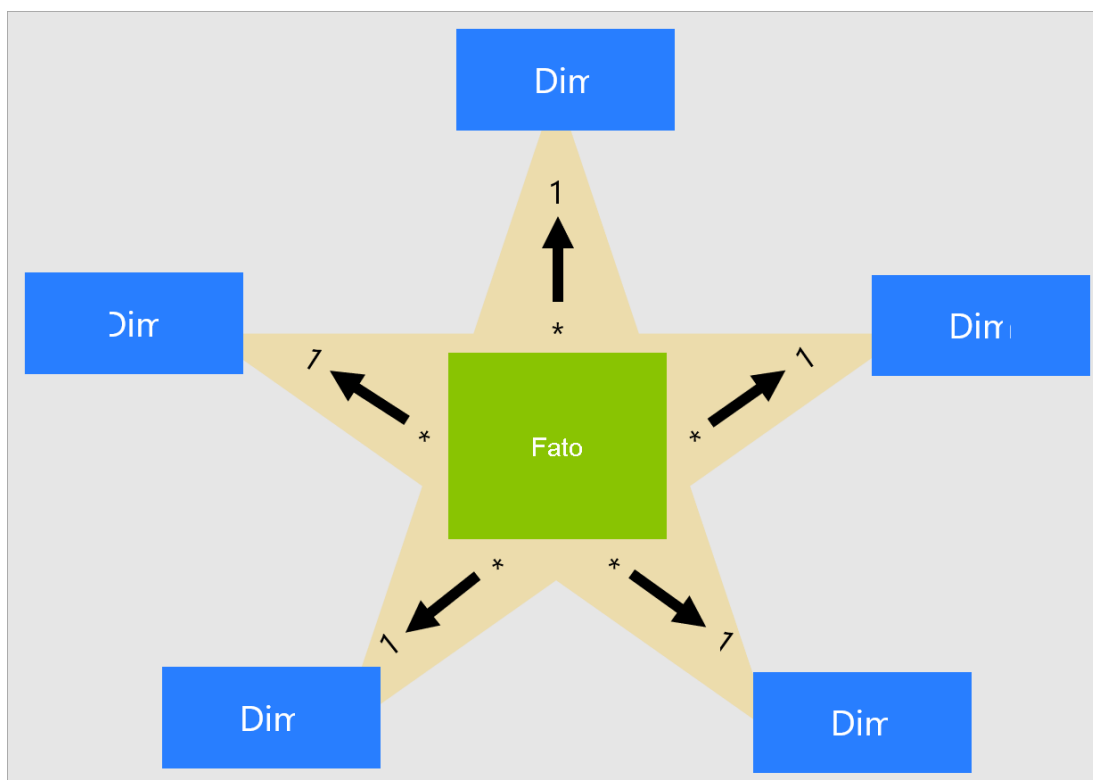


Figura 2: Ilustração esquema estrela. (Microsoft, 2023)

2.3.3 Ferramentas de visualização

Uma ferramenta de visualização para data warehouse é um software projetado para extrair, transformar e visualizar dados armazenados em um data warehouse. O objetivo principal dessas ferramentas é fornecer uma maneira intuitiva e interativa de explorar e compreender os dados do data warehouse, transformando-os em visualizações gráficas compreensíveis. Um exemplo é o Google Data Studio que é uma ferramenta gratuita de visualização de dados que permite criar painéis interativos e relatórios personalizados usando dados do de um data warehouse. (Oracle, 2023)

3 Proposta do Data Warehouse

A Companhia de Saneamento de Água e Esgoto é uma empresa estatal que fornece saneamento básico e que lida com uma grande quantidade de dados, incluindo informações de pagamento dos clientes, histórico de consumo, registro de ocorrências que podem ir desde vazamentos até possíveis fraudes.

Por se tratar de uma grande volume diferente de dados, a companhia tem várias fontes de dados dispersas, como leituras nas residências feita através dos TACEs (Técnico de Atendimento ao Cliente Externo) ou dados de atendimento online.

Um data warehouse ajudará a consolidar esses dados em um único local centralizado.

A empresa necessita ter análises avançadas para obter insights sobre os comentários do cliente, eficácia das campanhas de marketing e conscientização sobre o uso da água, médias de consumos e identificação de oportunidades de negócio. Junto a isso, as equipes que trabalham precisam de um rápido acesso às informações, para ter uma tomada rápida de decisão e também um atendimento eficaz junto ao cliente.

Antes de implementar um data warehouse em uma empresa nesse molde, é indispensável realizar uma avaliação detalhada de como esses dados são fornecidos e como são armazenados, também é importante analisar a infraestrutura da empresa e quais modificações serão necessárias para a implementação. Com isso em mente podemos olhar para alguns pontos importantes:

Requisitos de dados:

- As fontes de dados são os dados de leitura, histórico de consumo dos clientes, solicitações feitas pelos clientes via aplicativos, reclamações feitas através do telefones, registros de vazamentos e relatórios de atendimento. O que gera um volume de dados crescentes já que serão atualizados em tempo real.

Requisitos de análise:

- Para a Companhia é de suma importância que seja feita uma análise constante de todo o seu volume de dados para uma tomada de decisão mais lúcida, e uma maior eficiência na entrega de seus serviços de maneira geral, e para é isso é necessário que seja possível realizar pesquisas com um alto nível de detalhamento combinando as diversas fontes de dados que possui.

Requisitos de Segurança e privacidade:

- A Companhia tem acesso a dados de consumo de água e a dados referentes de pagamentos dos seus clientes, e apesar de também trabalhar com dados que não são confidenciais por se tratar de uma empresa pública é importante definir políticas de acesso e controle de dados para garantir a segurança e a privacidades das informações armazenadas no data warehouse.

Habilidades Técnicas:

- Identificar se a equipe possui habilidades técnicas em relação à implementação e manutenção do data warehouse, e caso não tenha identificar as competências necessárias e propor um treinamento adicional para a equipe

Infraestrutura e Recursos:

- Avaliar os recursos de hardware e software existentes na empresa e determinar se ela suporta o data warehouse e decidir se será necessário investimento em infraestrutura adicional como servidores, armazenamento, etc. Avaliar o orçamento disponível para o projeto de implementação.

3.1 Definição do esquema estrela

Como forma de implementação do data warehouse propomos o uso do esquema estrela, pois ao usá-lo facilmente para analisar os dados coletados das diversas fontes já citadas. A figura 4 mostra exemplifica a estrutura utilizada:

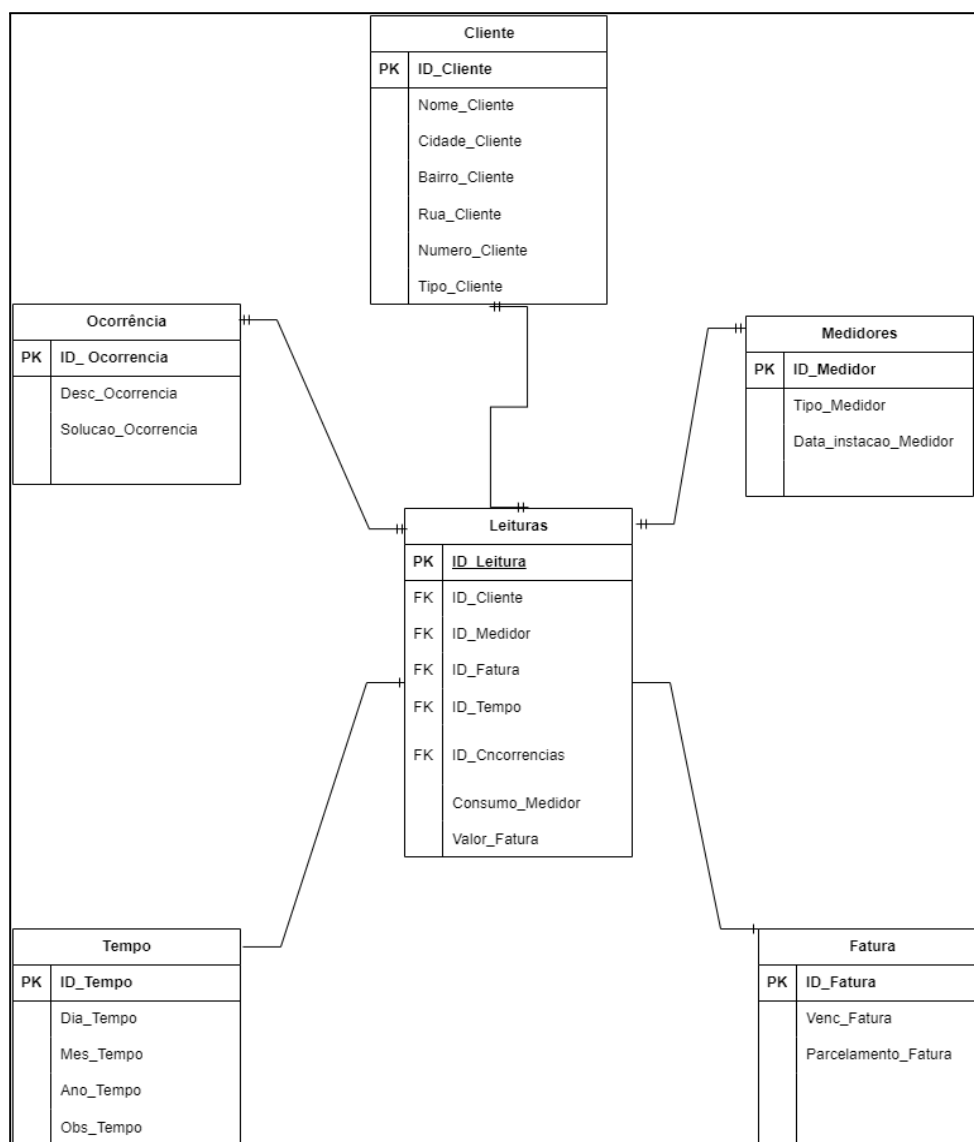


Figura 2: Ilustração esquema estrela. (Microsoft, 2023)

Tabela de fatos:

Leituras: Essa tabela principal que armazenará as informações sobre as leituras de consumo de água realizadas pela empresa. Ela pode incluir os seguintes campos:

- ID da leitura (chave primária)
- ID do medidor (chave estrangeira)
- ID do Cliente (chave estrangeira)
- ID da Ocorrência (chave estrangeira)
- ID Tempo (chave estrangeira)
- ID da Fatura (chave estrangeira)
- ID do cliente

Tabelas de dimensão:

Clientes: Essa tabela armazenará informações sobre os clientes da empresa de saneamento. Ela pode incluir os seguintes campos:

- ID do cliente (chave primária)
- Nome do cliente
- Endereço do cliente
- Tipo de cliente (residencial, comercial, industrial)

Medidores: Essa tabela armazenará informações sobre os medidores de água instalados nos clientes. Ela pode incluir os seguintes campos:

- ID do medidor (chave primária)
- Tipo de medidor (analógico, digital)
- Data de instalação do medidor
- Consumo (em m³)
- Localização do medidor (endereço do cliente)

Ocorrências: Esta tabela armazena informações de ocorrências registradas pelos clientes. Ela pode incluir os seguintes campos:

- ID da ocorrência (chave primária)
- Data da ocorrência
- Descrição
- Solucionado (sim ou não)

Tempo: Essa tabela armazenará informações sobre as datas e horários das leituras de consumo. Ela pode incluir os seguintes campos:

- ID do tempo (chave primária)
- Data
- Outras informações relacionadas ao tempo (opcional)

Fatura: Essa tabela armazena informações sobre as faturas. Ela pode incluir os seguintes campos:

- ID da fatura (chave primária)
- Valor da fatura
- Vencimento da fatura
- Parcelado (Sim ou não)

Com esse esquema estrela, podemos facilmente realizar análises relacionadas ao consumo de água e faturamento da empresa. Por exemplo, podemos determinar o consumo médio mensal de água, identificar os clientes que consomem mais água em um determinado período ou analisar o padrão de consumo ao longo do tempo, também podemos analisar casos de possível fraude e impedir que novos casos ocorram com base na análise detalhada dos dados.

4 CONSIDERAÇÕES FINAIS

Com o avanço constante da Tecnologia, a quantidade de dados continua a aumentar a cada dia, e junto surgem também as dificuldades das instituições de lidar com essas quantidades massivas de dados, e é justamente a capacidade de trabalhar, estudar e analisar esse grande volume de dados que chamamos de big data.

O big data tem sido cada vez mais utilizado por grandes instituições com a finalidade de obter percepções mais profundas sobre os mais variados setores, com o big data é possível tratar dados de diversas fontes e tirar deles informações realmente relevantes.

Para se trabalhar com big data existem vários modelos de banco de dados e nesse trabalho sugerimos a implementação de um Data Warehouse em uma companhia de saneamento de água e esgoto, que irá ajudar a tratar esses dados de uma maneira mais completa, trazendo assim resultados mais precisos.

Com o novo DW, a Companhia poderá realizar análises mais profundas e detalhadas, proporcionando insights valiosos sobre como comporta o cliente, a eficácia do seu marketing e de suas campanhas e do desempenho de seus negócios.

Com dados mais eficazes, a companhia saberá onde atuar, por exemplo, para se evitar o desperdício de água, tendo em vista que trará onde se está tendo essa grande “vazão” de água, no qual não está sendo consumida e paga, com isso, terá

uma tomada de decisão mais estratégica e ágil nos seus serviços. Ajudando também na redução do tempo gasto em serviços e pesquisas que acabavam não trazendo grandes resultados, pois agora todos esses dados já estão consumidos, preparados e analisados para poder “sair” para os serviços a serem executados nas ruas.

Com análises mais detalhadas, também poderá identificar emergências que a empresa necessita, buscando assim parcerias ou até mesmo privatizações de alguns de seus setores, para otimizar seu serviço a velocidade dele.

REFERÊNCIAS

SAS. O que é e sua importância: Disponível em:
https://www.sas.com/pt_br/insights/data-management/o-que-e-etl.html. Acesso em: 01/05/2023.

ORACLE. What is Big Data?: Disponível em:
<https://www.oracle.com/br/big-data/what-is-big-data/>. Acesso em: 08/04/2023.

Universidade Federal do Rio de Janeiro (UFRJ). Big Data e sua e suas influências sobre as estratégias das empresas: Disponível em:

https://www.gta.ufrj.br/grad/15_1/bigdata/vs.html#:~:text=Os%205%20V's%20do%20Big%20Data&text=A%20proposta%20de%20uma%20solução,a%20Veracidade%20e%20o%20Valor Acesso em: 08/04/2023

SAS. O que é o Hadoop?: Disponível em:
https://www.sas.com/pt_br/insights/big-data/hadoop.html Acesso em: 03/05/2023

DAVENPORT, T. Big Data no Trabalho: Derrubando Mitos e Descobrendo Oportunidades. 2014

MARR, B. Big Data: Using Smart Big Data Analytics and Metrics To Make Better Decisions And Improve Performance. 2015

ORACLE. What is Data Visualization? Disponível em:
<https://www.oracle.com/br/business-analytics/what-is-data-visualization/> Acesso em: 01/06/2023

AMAZON. What's Cloud Computing?. Disponível em:
<https://aws.amazon.com/pt/what-is-cloud-computing/> Acesso em: 01/05/2023

IBM. Apache Kafka. Disponível em:
<https://www.ibm.com/br-pt/topics/apache-kafka> Acesso em: 05/05/2023