Artigo

Pentaho BI - Conhecendo a Plataforma, Arquitetura e Infraestrutura

Veja nesse artigo como a suíte Pentaho BI pode ser utilizada para atender demandas que vão além de exibir simples relatórios.

A suíte Pentaho de Inteligência de Negócios é um conjunto de softwares livres que

Figura 1: Solução de BI com Pentaho Open BI Suite

Estão disponíveis componentes para execução de processos de ETL, que fazem carga de Data Warehouses, criação de relatórios pré-formatados e ad hoc, cubos **OLAP**, painéis de instrumentos (Dashboards) e garimpagem de dados (Data Mining). Todos esses recursos podem ser combinados e acionados sequencialmente para criação de soluções mais sofisticadas. Além disso, a plataforma executa todas as soluções de BI como serviços e, por isso, é possível prover acesso às soluções para sistemas externos, via web services, através de um mecanismo baseado em **SOAP/WSDL/UDDI**.

A suíte se divide em duas partes: a Pentaho BI Plataform propriamente dita, implementada na forma de um servidor web, e clientes de desenvolvimento, que criam conteúdo para a plataforma.

O Pentaho é um software patenteado nos **EUA**: os fundadores da empresa queriam desenvolver um pacote Java que pudesse ser usado para construir qualquer solução de **BI**. Eles queriam isso porque achavam que nenhuma ferramenta era flexível e poderosa o bastante para atender qualquer necessidade, de qualquer

Depois de desenvolver a primeira versão desse pacote, eles montaram uma demonstração de como usá-lo. Eles chamaram esse exemplo de "Pre-Configured Installation", ou **PCI**, mas ela fez tanto sucesso que a Pentaho foi forçada a adotá-la como produto e a evoluí-la. Essa trilha levou ao **Pentaho BI Server**, que contém o console de usuário (Pentaho User Console, **PUC**) e o console de administração (Pentaho Administration Console, **PAC**). E hoje em dia, quando falamos a plataforma Pentaho estamos nos referindo indistintamente tanto à esse "sub-produto" como a plataforma propriamente dita.

A Plataforma é uma aplicação em **JSP** que roda sobre um servidor de aplicações Java - até a versão 1.7GA o default era **JBoss**; a partir da 2.0GA passou a ser **Tomcat**. A plataforma se divide em duas partes:

- O Solution Engine e seus componentes, são responsáveis pela execução e controle das soluções. A base de seu funcionamento é uma máquina de workflow interna, que sequencia as chamadas de cada componente para o resultado desejado.
- O Portal, a porção do Pentaho visível ao cliente final. Através dele o cliente navega entre as soluções e aciona a execução de qualquer recurso, como um relatório ou dashboard.

A partir da versão 2.0 algumas funções foram movidas do Portal para o Administration Console, uma outra aplicação web e parte da suíte.

O BI Server oferece alguns serviços pré-configurados, como registro de soluções, controle de acesso, relatórios ad-hoc, agendamentos etc. Finalmente, a modularidade do portal permite que novos serviços sejam criados e

Servlets chamados Actions.

Todos os softwares da Suite Pentaho são programas Java e rodam em qualquer plataforma que tenha uma **JVM** padrão.

Business Intelligence Server, a encarnação mais famosa da plataforma, o **BI Server**, é uma aplicação Java Web, montada sobre um Tomcat, pré-configurada com vários recursos:

- Controle de acesso ao ambiente por usuário e senha;
- Controle de acesso aos objetos (pastas, relatórios, painéis etc.) baseado em usuários e papéis;
- Controle de acesso aos dados, que diz quem pode ver que registro, de que tabela;
- Relatórios AdHoc;
- Visualizador/Navegador **OLAP**;
- Relatórios pré-configurados (a priori);
- Agendador de relatórios;
- Execução de relatórios em background(plano de fundo);
- Envio de resultados por e-mail (bursting).

A interface visual do BI Server leva o nome de Pentaho User Console, ou **PUC**. Existe ainda uma outra aplicação, baseada em Jetty, que faz a administração da plataforma, com criação e gestão de usuários, papéis, fonte de dados e outros serviços como purga automática de conteúdo e controle de agendas públicas. Essa interface se chama Pentaho Administration Console, ou **PAC**.

Pentaho Report Designer (PRD), é o gerador de relatórios stand-alone da suíte Pentaho, representante da categoria tornada famosa pelo Crystal Reports. Ele pode conectar-se a qualquer fonte de dados para qual exista um driver **JDBC** e criar relatórios pixel perfect, exibindo não apenas lista de dados, mas também o resultado de fórmulas, subrelatórios, links, imagens, gráficos (pizza, barra, linha etc.), dentre outros recursos. A partir da versão 3.5, o PRD passou a oferecer parametrização de relatórios na própria ferramenta. O PRD pode ser usado sozinho, ou publicar os relatórios diretamente no BI Server, para posterior acesso via web.

O Pentaho Metadata Editor (PME), permite que o arquiteto da solução de BI agrupe os campos de tabelas que tenham alguma correlação, criando visões de negócios independentes, mesmo que campos de visões distintas residam em uma mesma tabela. Ele é totalmente visual, e pode mapear qualquer fonte de dados que possua um driver **JDBC**.

O Pentaho Schema Workbench (PSW), cria os cubos OLAP que serão exibidos na PUC. Ele tem uma interface visual para navegar entre as definições do cubo, permitindo criar métricas, dimensões e hierarquias.

Já o Pentaho Design Studio (PDS) é o ambiente de implementação de Soluções de BI, que cria Actions Sequences e as combina em soluções mais complexas. Ele fornece ao desenvolvedor acesso de baixo nível aos recursos do Pentaho e é um plugin para a IDE Eclipse.

O Pentaho Weka é um ambiente gráfico para Data Mining. Permite ao usuário criar e testar hipóteses contra as bases de dados.

fixo), planilhas Excel e base de dados **ODBC**. Ele é um ambiente gráfico no qual conexões com fontes de dados são estabelecidas e sequencia de passos executam a extração de dados, sua modificação e a carga desses em um destino. Pode integrar dados entre empresas e sistemas, substituindo a criação de camadas de programas para integração, por operações visuais.

A versão 4.0 implementa o conceito da Pentaho de AgileBI, no qual se combinam em uma interface a extração de dados, sua modelagem e relatórios. Modelos e relatórios podem ser publicados diretamente no BI Server. Essa integração permite que a equipe de BI gere resultados em dias ao invés de semanas.

A comunidade mundial Pentaho desenvolveu um número de outros recursos que podem ser adicionados ao Pentaho, notadamente ao BI Server.

Por definição, a Suite Pentaho acessa (lê/grava) qualquer base de dados para qual haja um driver JDBC. Além disso, em ambiente Windows, a Suite consegue ler de qualquer base de dados que tenha driver ODBC, através do driver JDBC para ODBC. A capacidade de gravação via ODBC não é uniforme (algumas bases dispõem, outras não). Na **Figura 2** temos uma relação dos bancos empacotados no Pentaho.

Banco de Dados Empacotados no Pentaho Data Integration	
Apache Derby	MonetDB
AS/400	MS Access (OBDC)
Borland Interbase	MS SQL Server
DB2	MySQL
dBase III, IV ou 5 (ODBC)	Neoview
ExtenDB	Netezza
Firebird SQL	Oracle
Greenplum	Oracle RDB
Gupta SQL Base	Palo MOLAP Server (via plugin.)
H2	PostgreSQL
Hypersonic (ex-HSQLDB)	Remedy Action Request System (ODBC)
Infobright	SAP R/3 System (via plugin.)
Informix	SQLite
Ingres	Sybase
Intersystems Cache	SybaseIQ
KingbaseES	Teradata
LucidDB	UniVerse database
MaxDB (SAP DB)	Vertica

Figura 2: Lista de Banco de Dados PDI

A criação de soluções de BI com a Plataforma Pentaho obedece a um fluxo simples, conforme mostrado na **Figura 3**:

1. Os clientes, PDI, PRD, PSW, PDS, PDA, criam os artefatos da solução;
2. Esses artefatos são publicados no BI Server;
3. Os usuários acessam o BI Server para executar as soluções. Uma solução de BI precisa de fontes de dados confiáveis e de alguma interface para seu cliente explorá-los. Algum tempo depois, a exploração eventualmente amadurece em um processo, que pode ser automatizado, gerando valor para empresa.

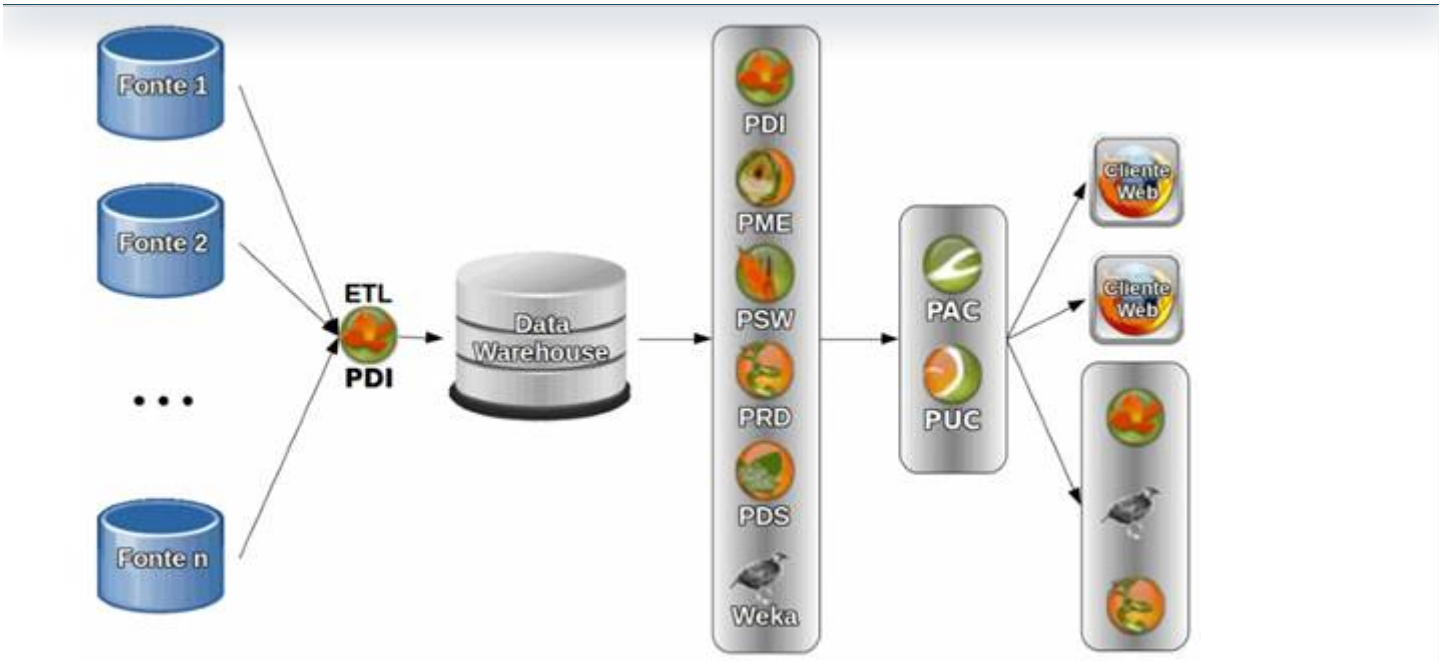


Figura 3: Processo de criação de Solução de BI Padrão com Pentaho.

Os passos destacados correspondem à:

1. Criação de Data Warehouse, Data Mart ou dump do banco de dados com o Pentaho Data Integration, a partir de fontes de dados que podem ser bancos relacionais, serviços de rede, páginas web e fontes desestruturadas (como e-mail e documentos texto), além de arquivos planos (CSV, Excel, Etc.).
2. Criação das soluções iniciais para exploração do repositório de dados: Cubos OLAP, relatórios (com ou sem parâmetros), WAQR. Todos os clientes de desenvolvimento podem ser usados.
3. Entrega da solução com BI Server, com controle de acesso via web por seus clientes. Alguns clientes podem ter demandas especiais e optar por usar algum dos clientes de desenvolvimento, como PDI, Weka ou Report Designer para atendê-las.

Toda solução de BI sempre tem três partes:

- Data Warehouse: como não existe BI sem DW, essa é uma peça indispensável em projetos de BI. Quando falamos de DW no contexto da arquitetura de BI invariavelmente estamos nos referindo ao servidor de banco de dados - hardware e software - que vai cumprir a função de armazém de dados para a solução de BI da empresa. Para definir esse componente é importante conhecer o volume de dados que será carregado inicialmente, a que velocidade (em bytes ou registros por mês) ele vai crescer, quanto usuários poderão consultá-lo e quantas estrelas ele vai ter. Normalmente nenhuma dessas informações é conhecida a priori, de modo que podemos quando muito fazer estimativas mais ou menos calibradas;
- Servidor de ETL: se DW na infraestrutura significa a máquina do banco de dados, ETL nesse contexto significa a máquina que vai executar o processo de extração, transformação e carga das fontes de dados para dentro do DW. De novo, como não há BI sem DW, não pode haver BI sem servidor de ETL porque a carga de um DW se dá por esse processo. Portanto, servidores de ETL também são indispensáveis na arquitetura de uma solução de BI;
- Servidor de Exploração de Dados: uma vez que os dados estejam disponíveis no DW, os usuários começam a acessá-los e a explorá-los para resolver suas diversas necessidades: medir o desempenho da empresa, responder as perguntas estratégicas, táticas e até mesmo operacionais, planejar e avaliar o resultado das ações e um inimaginável sem números de usos.

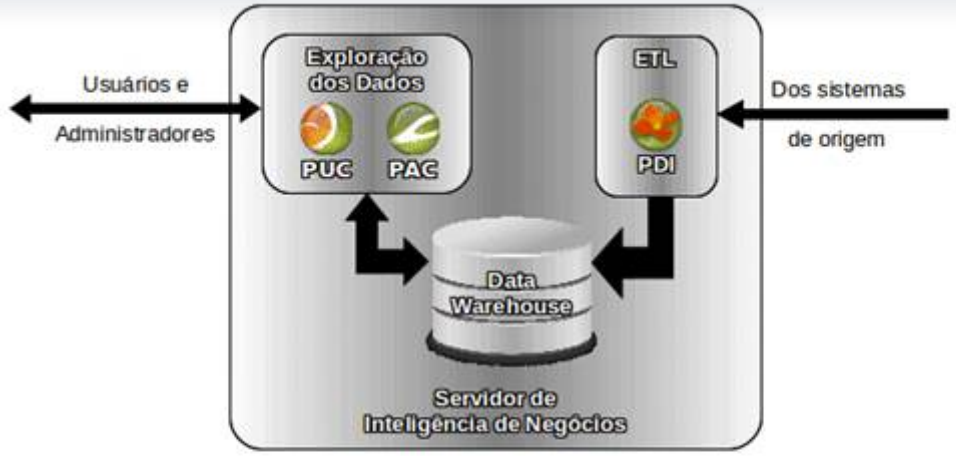
Alguns projetos de BI, como os que envolvem Data Mining, consomem os dados na forma de arquivos extraídos do DW especialmente para essas necessidades, e

para realizar consultas SQL que populam planilhas Excel simplesmente matam o interesse, não pela falta de versatilidade, mas pela falta de usabilidade e de prazer em trabalhar com esses dados.

Por isso todo projeto de BI que se preze oferece aos usuários finais, seus clientes, um programa que dê essa interface gráfica. Até meados da década de 2000 ainda existiam softwares stand-alone, que eram instalados na estação de cada usuário. Mas uma tendência nascida na década anterior estava atingindo a maturidade: interfaces para DW em ambientes web, ou cliente-servidor como eram chamados.

Esse é o terceiro componente indispensável de uma solução de BI: um software que ofereça ao cliente uma poderosa e agradável interface gráfica para exploração de dados do DW.

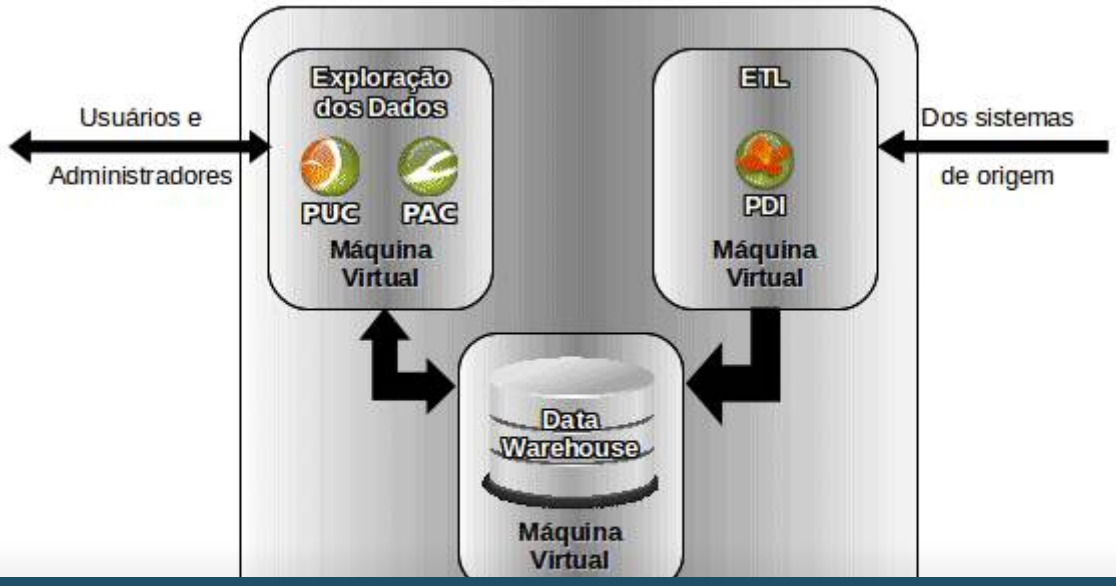
- Tudo-Em-Um Hardcore, ideal para os projetos pilotos, pequenas empresas ou o início de projetos ágeis. A combinação de todos os servidores em uma só máquina física e lógica é uma boa opção porque oferece menor complexidade e maior facilidade de gestão, preço reduzido e menor consumo de mão-de-obra especializada. É um ótimo ambiente para experimentações ou para projetos departamentais, conforme mostrado na **Figura 4**.



CPU	2GHz, dual-core
RAM	12GB
HD	100GB
Rede	1Gbit

Figura 4: Esquema tudo-em-um, hardcore.

- · Tudo-Em-Um Softcore, a empresa que sabe que sua necessidade de BI vai crescer pode usar uma variação do modelo anterior: uma única máquina, mais parruda que a média, mas com três máquinas virtuais, conforme mostrada na **Figura 5**.



Quando a necessidade de poder de processamento, memória, disco ou rede aumentar, a virtualização dá mais opções de reestruturação. Por exemplo, separar os servidores por demanda em uma fazenda de servidores virtualizados.

Virtualização, com o barateamento do hardware de prateleira, torna-se possível criar ambientes virtualizados cada vez mais poderosos em máquinas cada vez mais baratas. A extrapolação do modelo Tudo-Em-Um Softcore leva a uma estrutura de nuvem, conforme mostrado na **Figura 6**.

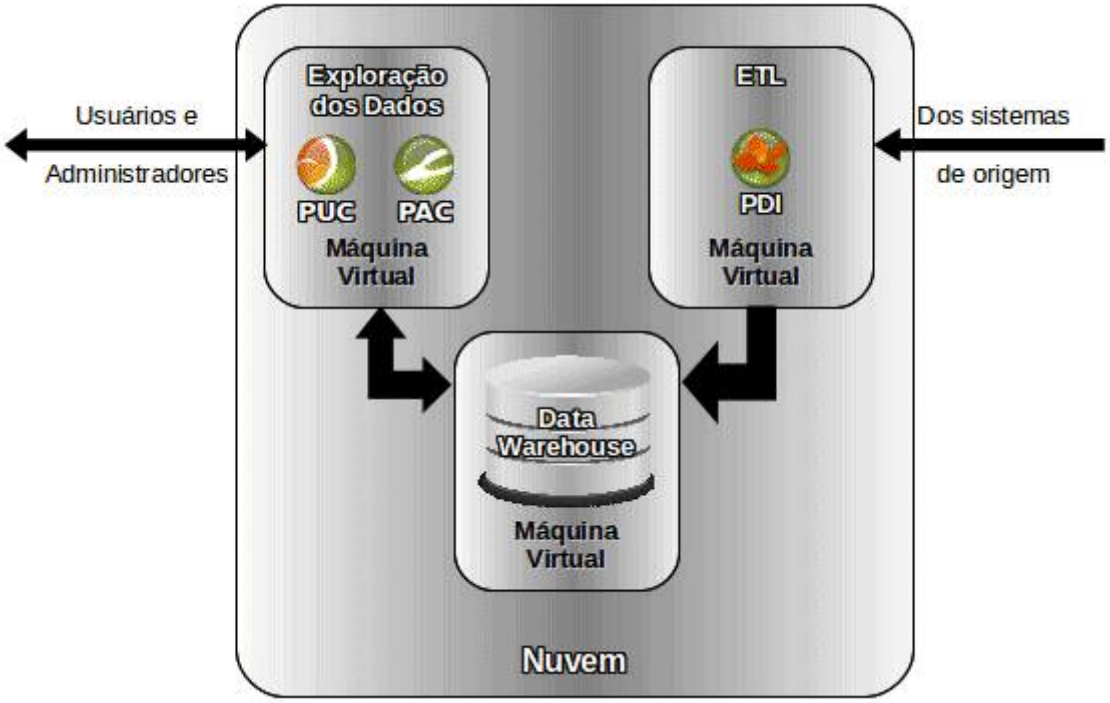


Figura 6: Virtualização total (nuvem).

A popularização e o barateamento de software de clusterização dinâmica (cloud computing) - permite que a empresa invista em um ambiente inicial e o expanda a medida que a demanda crescer.

A maior vantagem dessa arquitetura é a capacidade de expansão praticamente

Servidores Independentes, organizações que conseguem estimar com precisão o crescimento da demanda sobre o ambiente de BI podem optar por um esquema no qual todos os servidores são reais e separados, conforme mostrado na **Figura 7**.

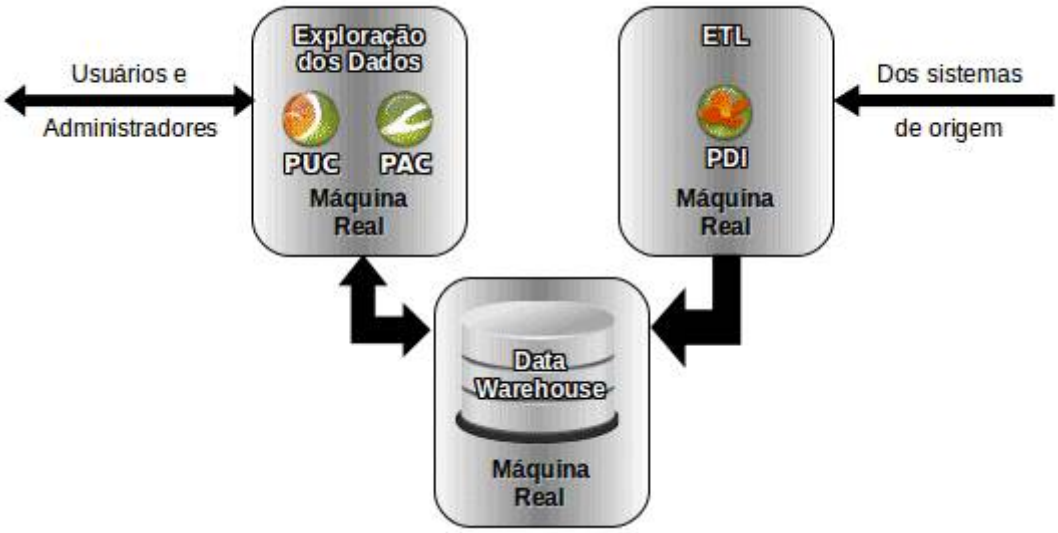


Figura 7: Servidores reais e independentes

A maior vantagem é a economia decorrente do autoconhecimento. Como a empresa conhece bem a própria demanda ela pode planejar a evolução de cada ambiente e com isso espaçar mais as compras de hardware. Outro benefício colhido é a economia de gerenciar apenas três servidores físicos. Gerir esses servidores é mais barato que gerir uma nuvem pois não requerem a administração da arquitetura de nuvem além das instâncias de banco de dados e servidores, especialmente do ponto de vista de mão-de-obra. Finalmente, é possível crescer memória e CPU da máquina que se tornar um gargalo com alguns upgrades relativamente baratos, antes de trocar por máquinas mais potentes.

Figura 8: Servidores DW e ETL combinados.

O processo de ETL também pode ficar dentro do servidor de exploração, conforme mostrado na **Figura 9**.

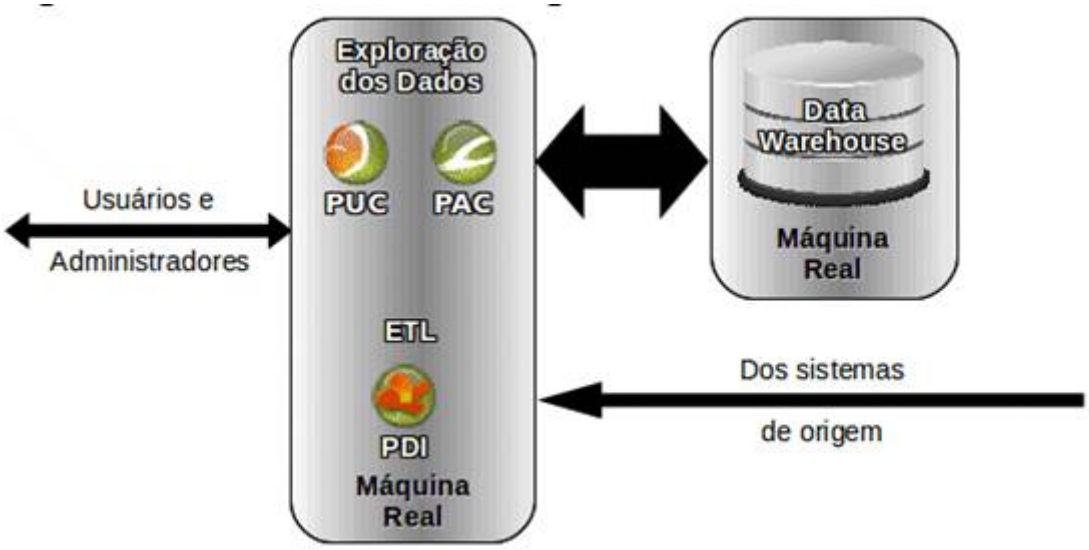


Figura 9: Processo ETL roda dentro do servidor de exploração.

Essa combinação é melhor que a anterior porque dá máquinas inteiras dedicadas a cada parte do processo, sem comprometer-se entre si. Na anterior, CPU, memória e disco usado pelo processo de ETL era subtraído do banco de dados, e vice-versa.

Sua necessidade de hardware vai depender diretamente do ambiente que você deseja implantar. Como linhas gerais, e em ordem decrescente de importância, busque:

- Maior capacidade de expansão de memória;
- Maior capacidade de rede (para ambientes separados);
- HDs com maior vazão;
- Maior capacidade de CPU;
- Isso porquê:
  - O maior gargalo na expansão de usuários é espaço para todas as sessões simultâneas;
  - O maior gargalo para consultas simultâneas é a troca de dados entre o servidor DW e o de exploração;
  - O maior gargalo à troca de dados é a velocidade de acesso aos dados em disco;
- E só depois desses gargalos resolvidos é que vai adiantar aumentar o poder de processamento, pois sem dados prontamente disponíveis para todos os usuários não adianta nada ter CPU potente.

Até a próxima! Um abraço.

Referências:

THOMSEN, ERICK. OLAP Solutions. 1a. Ed. EUA: Wiley Publishing, 1997.

SCHEPS, SWAIN. Business Intelligence for Dummies. 1a. Ed. EUA: Wiley Publishing, 2008.

Pentaho na prática (Fábio, Caio & Cesar) ISBN: 978-85-915459-0-2