UNIVERSIDADE FEDERAL DE SÃO CARLOS

CAMPUS SOROCABA

Orientações para Projeto

Disciplinas: **Novas Tecnologias em Banco de Dados (BCCS)**

Banco de Dados (PPGCCS)

Período: **1/2014**

Objetivos:

O foco deste projeto será a criação de um protótipo de *data warehouse* em torno de um tópico específico, para:

1. aprender e descrever os componentes básicos de um DW;
2. projetar um DW com base na definição de requisitos de um problema de negócio;
3. criar um DW protótipo usando princípios discutidos em sala de aula.

Formação da Equipe:

Deverão ser formados grupos de até 4 alunos.

Descrição das etapas do projeto:

O projeto consta de duas etapas de desenvolvimento. Elas são:

1. Modelagem, fonte de dados e ETL:
   - *Levantamento de requisitos*: os usuários e seus requisitos de negócios impactam em quase todas as decisões feitas durante a implementação do DW. Portanto, essa etapa consiste daquelas atividades necessárias para obter uma compreensão completa da área de negócio específica.
   - *Modelagem*: envolve o projeto lógico do assunto de negócio, a análise detalhada das fontes de dados e as regras de transformação necessárias.
   - *Projeto físico*: mapeamento da modelagem lógica em um banco de dados físico.
   - *ETL*: envolve a concepção e desenvolvimento dos processos que carregam dados dos sistemas operacionais para o *data warehouse*. O DW deve atingir, no mínimo, um 1Gb de armazenamento.
2. OLAP + visualização:
   - *Finalização da ETL.*
   - Os *templates* de *front-end* são projetados durante esta fase. Crie relatórios padronizados e outros, de acordo com os requisitos levantados.

Além das fases acima, no projeto de um DW inclui-se uma fase de Implantação, que envolve a implantação real do ambiente de DW para a comunidade empresarial (NÃO inclusa no projeto).

Requisitos Obrigatórios:

- O modelo deve ter no mínimo **três** dimensões, **três** hierarquias (não precisa ser uma para cada tabela) e **uma** medida na tabela fato. Será aceito qualquer tipo de modelo estudado em aula.
- Devem existir no mínimo **duas** fontes de dados diferentes (diferente modelo lógico de dados), indistintamente de ser interna ou externa.
- A utilização de recursos avançados vistos em aula, tais como aspectos de modelagem avançada e índices para DW, serão valorizados na avaliação do projeto. É obrigatório a utilização de pelo menos um desses recursos (*Surrogated key* não conta como recurso avançado).
- **Uma** transformação de dados fora as consolidações de dados na fase de ETL.

Tecnologias:

A IMPLEMENTAÇÃO DO PROJETO:

- **ROLAP EM QUALQUER SGBD RELACIONAL, PREFERENCIALMENTE POSTGRESQL OU MS SQL SERVER**
- **QUALQUER FERRAMENTA DE BI, TANTO DE ETL COMO SERVIDOR OLAP OU DE VISUALIZAÇÃO**

Algumas tecnologias livres ou comerciais (algumas com período *trial*), que podem ser usadas no desenvolvimento do projeto:

- Kettle (ETL Pentaho)
- Mondrian (servidor OLAP)
- Pentaho BI Suite
- Tableau Software (BI)
- Microsoft Analysis Services

Avaliação:

CADA GRUPO DE ALUNOS FICARÁ ENCARREGADO DE DESENVOLVER O PROJETO MAIS CRIATIVO, DIFERENCIADO E ATRATIVO QUE OS OUTROS GRUPOS.

Na medida que seu projeto é o foco principal deste curso, todos os membros do grupo deverão ser capazes de falar de forma inteligente sobre o foco do projeto de seu grupo e o tema que está cobrindo a qualquer momento após o início do projeto. Se não o fizer, quando perguntado, poderá impactar negativamente a sua avaliação.

O projeto tem duas formas de avaliação complementares: execução e apresentação. O projeto será composto de uma fase intermediária, que permitirá um acompanhamento passo a passo da sua evolução.

- uma apresentação individual - **40**% da nota da fase
- a execução (código fonte correspondente à fase) - **60**% da nota da fase

A nota do projeto prático é calculada como:

```
FI * 40% + FF * 60%
```

Nas fases especificadas, deverá ser depositado no sistema de apoio *Moodle*, antes da data/horário marcados no cronograma, a apresentação com a descrição dos passos desenvolvidos + modelo/código fonte da execução.

ATENÇÃO:

- Alunos que não entregarem o projeto e/ou não realizarem sua apresentação final, por motivo justificado, poderão entregá-lo com prazo de uma semana após sua data limite inicial, com perda de 20% da nota. Excedido esse prazo, a nota desta fase de avaliação não mais será contemplada (zero), inclusive, quando o aluno apenas não realizou sua apresentação final.
- Haverá desconto na nota por dia de atraso na entrega das partes.
- Será avaliado para os diferentes grupos: a aplicação de conceitos, solução das consultas, corretude da informação, etc.
- Os erros detectados antes ou durante a fase intermediária devem ser corrigidos na fase seguinte, com maior peso de penalização na fase seguinte caso não forem corrigidos.
- **Não serão aceitas soluções idênticas de diferentes grupos (serão invalidadas na correção do projeto).**

Datasets:

A seguir são listados alguns *datasets* que podem servir para construir as fontes de dados do seu trabalho.

- Dados abertos do governo federal: http://dados.gov.br/
- INPE – dados meteorológicos e ambientais, dentre (http://satelite.cptec.inpe.br/PCD/)
- IBGE: censos e outros
- ANP – dados sobre petróleo, royalties do petróleo, etc.
- Machine Learning Repository http://archive.ics.uci.edu/ml/datasets.html
- Data Catalog.gov http://data.dc.gov/
- UK National Atmospheric Emissions Inventory datasets http://www.naei.org.uk/data_warehouse.php
- Prefeituras – dados de prestação de contas, portal da transparência

Ferramentas de geração de dados sintéticos:

Para atingir o volume de armazenamento mínimo, podem ser usadas ferramentas de geração de dados sintéticos. Por exemplo:

- *Parallel Data Generation Framework* (PDGF) - http://www.paralleldatageneration.org/
- *Generate data* - http://www.generatedata.com/
- *SQL Data Generator* (trial) - http://www.red-gate.com/products/sql-development/sql-data-generator/
- http://www.databasetestdata.com/
- *Datanamic Data Generator MultiDB* (trial) - http://www.datanamic.com/datagenerator/index.html
- *json-generator*: http://www.json-generator.com/

Cronograma:

Os alunos deverão desenvolver o projeto respeitando as fases, entregas e datas descritas a seguir.

Tabela 1: Cronograma de Execução do Projeto

Data	Fase	Descrição
05/09	**Formação dos grupos**	
19/09	**Definição dos temas**	
24/10	**BD-Intermediaria-1:**	**Modelagem e ETL**
05/12	**Entrega Final** **Apresentação Final**	**Entrega do código e apresentação final**

Entrega parcial ou intermediária: Fase 1 (com **80%** da implementação da fase de ETL)

Entrega final: corresponde ao projeto completo

As fases foram descritas em "Descrição das etapas do projeto".

Apresentações:

- Cada grupo deverá fazer uma apresentação oral de 20 minutos do andamento da fase do projeto em questão, seguida de arguição (5 minutos).
- O foco das apresentações é o acompanhamento dos passos desenvolvidos para atingir os resultados da fase.
- **Na apresentação intermediária**, esperasse como resultados a descrição do objeto de negócio, o modelo de DW, incluindo as principais consultas, a seleção das fontes de dados já implementadas e o início da implementação da fase de ETL. Nessa fase é importante deixar claro quem é o gestor do projeto, quais são os objetivos e principais consultas da análise multidimensional. O 80% da ETL inclui, no mínimo, a escolha das fontes de dados e o mapeamento entre as fontes e o modelo multidimensional, de forma a mostrar a origem de todos os dados do modelo e que as análises previstas são possíveis de serem executadas. Cada aspecto da implementação da ETL deve ser ao menos testada para um conjunto de dados.
- **Na apresentação final**, deve ser apresentado o sistema de suporte a decisão funcional (a aplicação rodando). Também, deve ser especificado se os objetivos foram atingidos, as alterações feitas referente à fase anterior, as dificuldades e o que não está funcionando. Mostrar o 1GB de dados gerados.
- **Para que não aconteçam erros na hora, que serão penalizados, por favor, testem o sistema funcionando no computador onde irá acontecer a apresentação. Treinem, também, a apresentação, para que a mesma aconteça de forma fluida e integrada.**
- Todos os membros do grupo devem apresentar. Aquele que não apresentar ficará sem nota na apresentação, porém, no caso da apresentação final, a nota da fase será ZERO.
- Se durante a apresentação é detectado que algum aluno não cumpriu a tarefa de participar na elaboração do projeto, a nota da fase será ZERO para execução, relatório e apresentação.