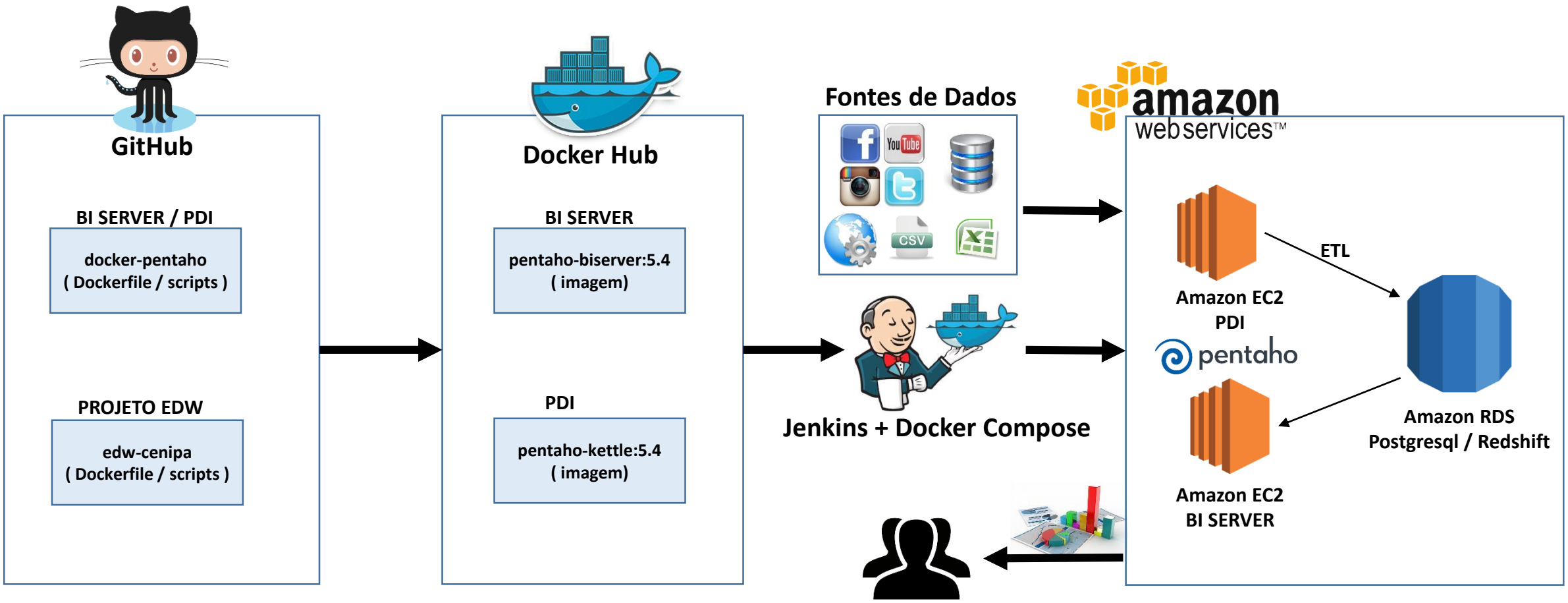
Construindo um data warehouse com Pentaho e Docker

Estudo de Caso: CENIPA
http://dados.gov.br/dataset/ocorrencias-aeronauticas-da-aviacao-civil-brasileira

Fonte do Projeto
https://github.com/wmarinho/edw_cenipa

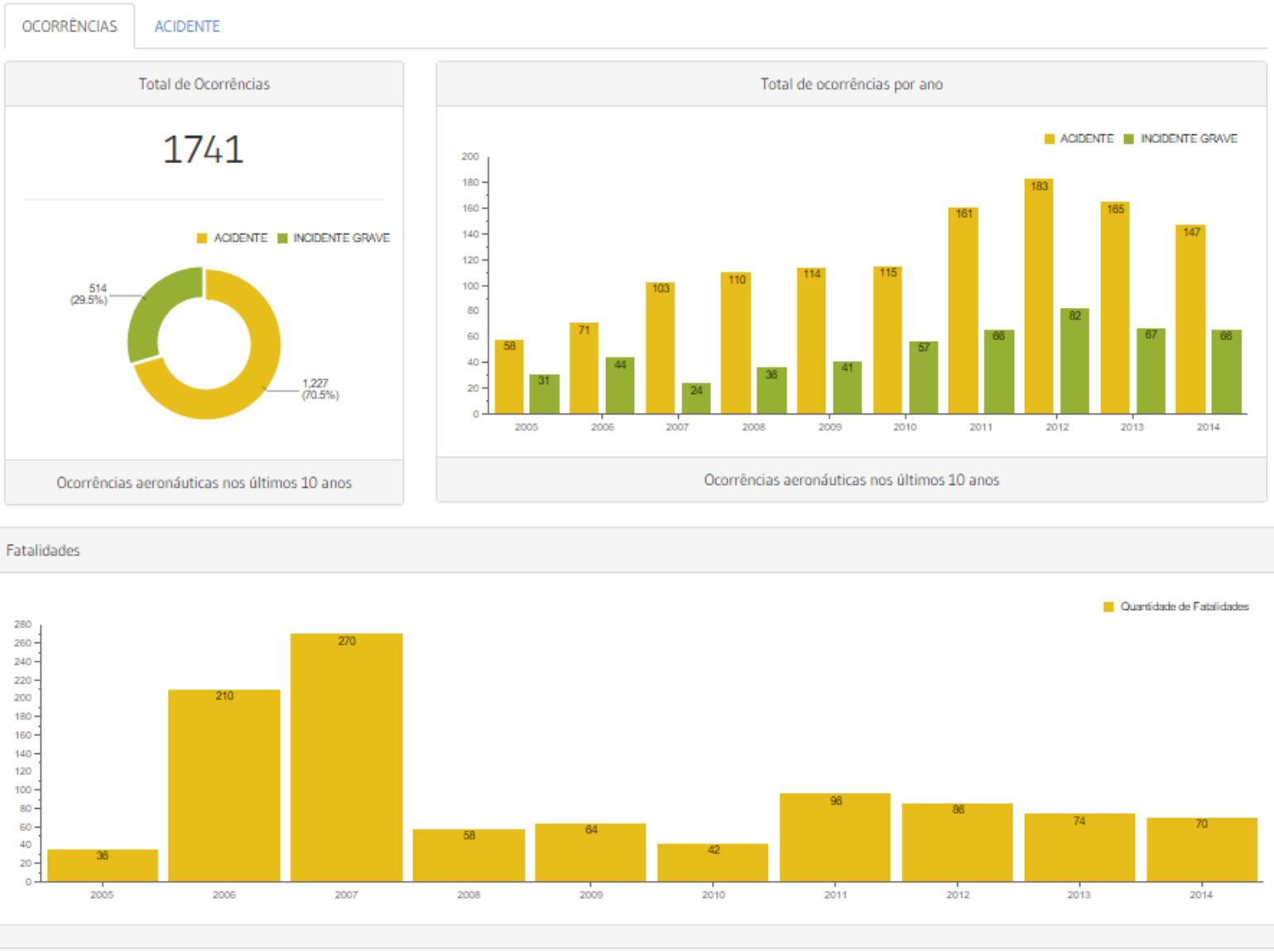
Wellington Marinho
wpmarinho@globo.com

Arquitetura

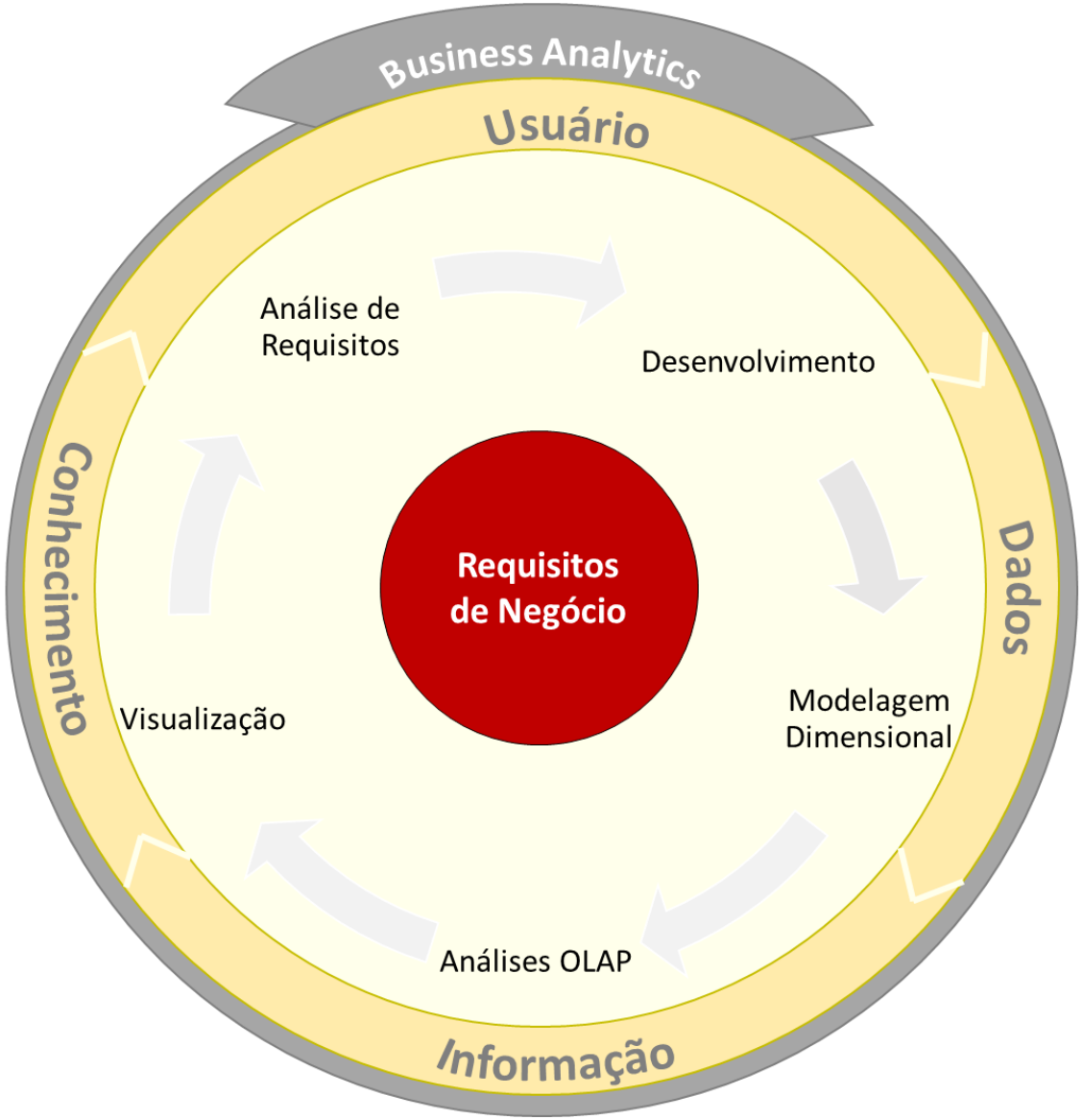


Dashboards – Ocorrências aeronáuticas

http://localhost/pentaho/plugin/cenipa/api/ocorrencias



Business Analytics



Estudo de Caso - EDW CENIPA

EDW CENIPA é um projeto opensource, criado para prover análises dinâmicas de ocorrências aeronáuticas, ocorridas na aviação civil brasileira. O projeto utiliza técnicas e ferramentas de BI, explorando tecnologias inovadoras e de baixo custo. Historicamente, plataformas de Business Intelligence são caras e inviáveis para pequenos projetos. Esses projetos exigem qualificação especializada e custos altos de desenvolvimento. Este trabalho tem a pretensão de quebrar um pouco esta barreira. O que não significa pouca dedicação, empenho e esforço.

Todas as análises têm como base os dados abertos fornecidos pelo CENIPA, com histórico de ocorrências dos últimos 10 anos:

- http://dados.gov.br/dataset/ocorrencias-aeronauticas-da-aviacao-civil-brasileira

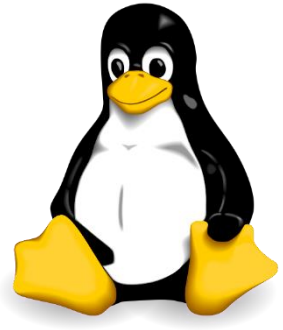
Os gráficos foram inspirados no relatório disponibilizado no link:

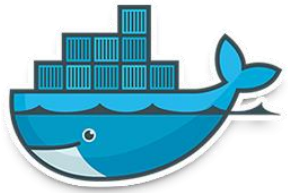
- http://www.cenipa.aer.mil.br/cenipa/index.php/estatisticas/estatisticas/panorama.

Recursos

Seguem alguns serviços, ferramentas e plataformas que foram utilizados para desenvolver e implantar o projeto:

- Amazon Web Services - https://aws.amazon.com/ - Serviços de infraestrutura de nuvem
- Sistema Operacional Linux - CentOS 6 / Ubuntu 14
- GitHub - https://github.com/ - Serviço de Web Hosting Compartilhado para projetos que usam o controle de versionamento
- Docker - https://www.docker.com/ - Plataforma aberta para construir e rodar aplicações distribuídas.
- Pentaho - http://www.pentaho.com/ e http://community.pentaho.com/ - Plataforma open source de Big Data, Data Integration e Business Analytics





Requisitos

- Sistema Operacional com 2GB de RAM e 5GB de espaço em disco
- Docker v1.7.1
  - CentOS: https://docs.docker.com/installation/centos/
  - Ubuntu: https://docs.docker.com/installation/ubuntulinux/
  - Mac : https://docs.docker.com/installation/mac/
- Docker Compose v1.4.2 - https://docs.docker.com/compose/install/

Instalação rápida no Amazon Linux AMI

```
$ yum update -y
$ yum install -y docker
$ service docker start
$ usermod -a -G docker ec2-user
$ yum install -y git
$ pip install -U docker-compose
$ PATH=$PATH:/usr/local/bin
```

Comandos básicos

```
$ docker info
$ docker --help
$ docker COMMAND --help
$ docker run --rm -it busybox echo "Olá, esse é meu primeiro container“
$ docker ps
$ docker images
$ docker build –t repositorio/imagem:tag .
```

Criar um arquivo Dockerfile

```
FROM busybox

CMD ["echo", "Olá, esse é meu primeiro container"]
```

Construir umaa imagem

```
$ docker build -t teste/myimage .
```

Criar um container

```
$ docker run --rm teste/myimage
```

Pentaho + Docker – Criando arquivo Dockerfile

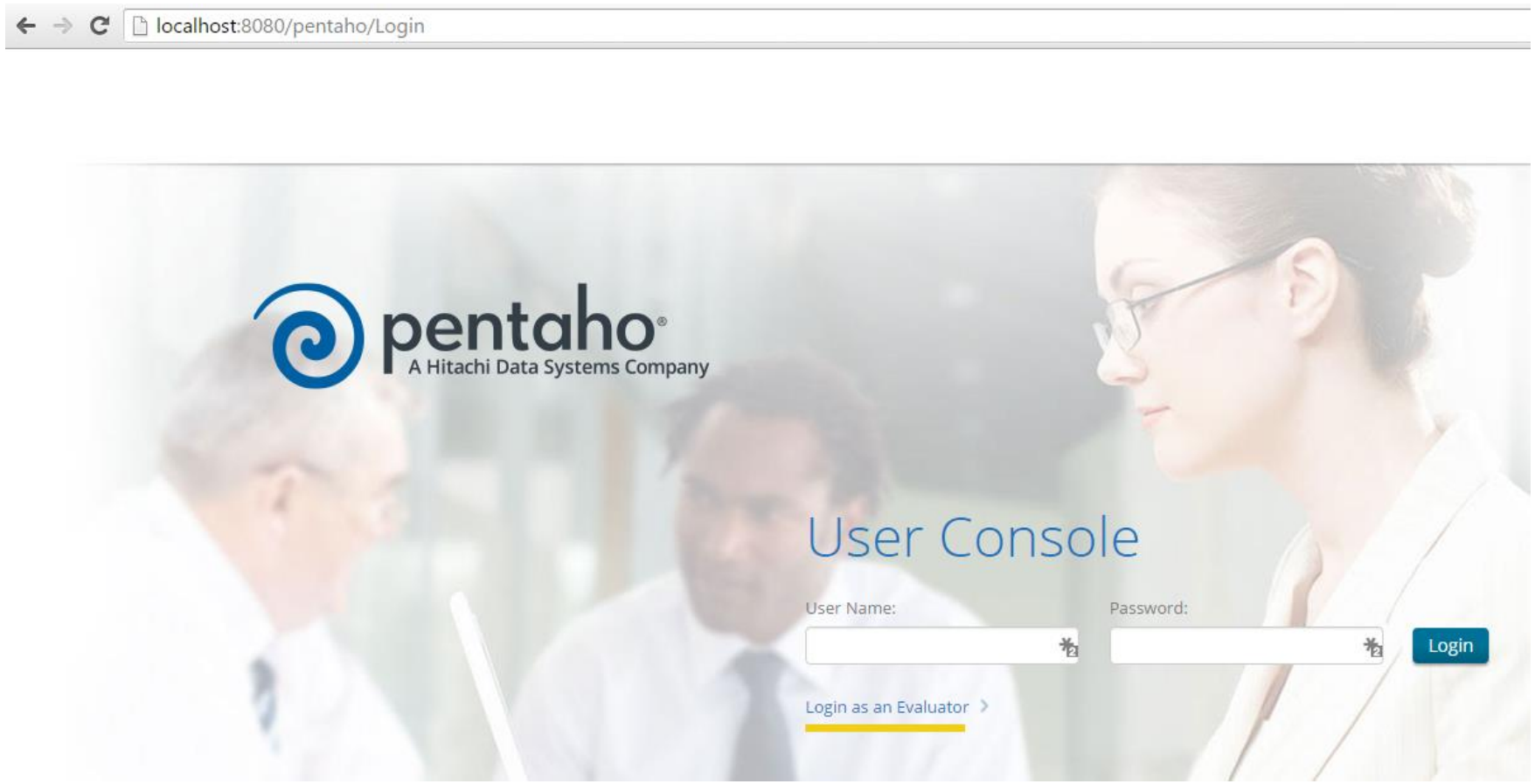
```
FROM java:7

MAINTAINER Wellington Marinho wpmarinho@globo.com

# Init ENV
ENV BISERVER_VERSION 5.4
ENV BISERVER_TAG 5.4.0.1-130

ENV PENTAHO_HOME /opt/pentaho

# Apply JAVA_HOME
RUN . /etc/environment
ENV PENTAHO_JAVA_HOME $JAVA_HOME
ENV PENTAHO_JAVA_HOME /usr/lib/jvm/java-1.7.0-openjdk-amd64
ENV JAVA_HOME /usr/lib/jvm/java-1.7.0-openjdk-amd64

# Install Dependences
RUN apt-get update; apt-get install zip -y; \
    apt-get install wget unzip git -y; \
    apt-get clean && rm -rf /var/lib/apt/lists/* /tmp/* /var/tmp/*;

RUN mkdir ${PENTAHO_HOME};

# Download Pentaho BI Server
RUN /usr/bin/wget --progress=dot:giga http://downloads.sourceforge.net/project/pentaho/Business%20Intelligence%20Server/${BISERVER_VERSION}/biserver-ce-${BISERVER_TAG}.zip
-O /tmp/biserver-ce-${BISERVER_TAG}.zip; \
    /usr/bin/unzip -q /tmp/biserver-ce-${BISERVER_TAG}.zip -d  $PENTAHO_HOME; \
    rm -f /tmp/biserver-ce-${BISERVER_TAG}.zip $PENTAHO_HOME/biserver-ce/promptuser.sh; \
    sed -i -e 's/\(exec ".*"\) start/\1 run/' $PENTAHO_HOME/biserver-ce/tomcat/bin/startup.sh; \
    chmod +x $PENTAHO_HOME/biserver-ce/start-pentaho.sh

RUN useradd -s /bin/bash -d ${PENTAHO_HOME} pentaho; chown -R pentaho:pentaho ${PENTAHO_HOME};

#Always non-root user
USER pentaho
WORKDIR /opt/pentaho

EXPOSE 8080
CMD ["sh", "/opt/pentaho/biserver-ce/start-pentaho.sh"]
```

Pentaho BI Server

Construir a imagem e rodar um containter

```
$ docker build -t pentaho/biserver:5.4 .
$ docker run --rm -p 8080:8080 -it pentaho/biserver:5.4
```

Abrir o navegador e abrir o Pentaho BI Server



Projeto EDW CENIPA

Instalação do Projeto

```
$ wget -O - https://raw.githubusercontent.com/wmarinho/edw_cenipa/master/easy_install | sh
```

A instalação pode levar mais de 30 minutos, dependo da configuração do servidor e da largura de banda da Internet. A instalação completa é de aproximadamente 3GB.

Verificar execução dos containers

```
$ docker ps
```

O projeto possui 3 containers:

- edwcenipa_db_1 - Container com Banco de Dados PostgreSQL
- edwcenipa_pdi_1 - Container com instlação do Pentaho Data Integrator (Kettle) para download e carga dados para o DW
- edwcenipa_biserver_1 - Container com instalação do Pentaho Business Analytics (BI Server)

Verificar logs

```
$ docker logs -f edwcenipa_pdi_1
$ docker logs -f edwcenipa_biserver_1
```

Docker Compose

docker-composse.yml - Orquestra a execução de múltiplos containers

```
 pdi:
  image: image_cenipa/pdi
  links:
     - biserver:edw_biserver
  volumes:
    - /data/stage:/tmp/stage
  environment:
    - PGHOST=172.17.42.1
    - PGUSER=pgadmin
    - PGPASSWORD=pgadmin.
    - PENTAHO_DI_JAVA_OPTIONS=-Xmx2014m -XX:MaxPermSize=256m
biserver:
  image: image_cenipa/biserver
  ports:
    - "80:8080"
  links:
    - db:edw_db
  environment:
    - PGUSER=pgadmin
    - PGPASSWORD=pgadmin.
    - INSTALL_PLUGIN=saiku
    - CUSTOM_LAYOUT=y
db:
  image: wmarinho/postgresql:9.3
  ports:
    - "5432:5432"
```

Pentaho + Docker + Amazon

Com o comando abaixo e as devidas credenciais de acesso, é possível subir o ambiente na Amazon em menos de 10 minutos. LEMBRE-SE de substituir as variáveis antes de executar o comando (verificar os parâmetros no AWS console).

Essa é uma configuração adequada para este projeto, a um custo aproximado de US$ 80,00/mês (http://calculator.s3.amazonaws.com/index.html)

```
$ SUBNET_ID=
$ SGROUP_IDS=
$ KEY_NAME=
$ aws ec2 run-instances \
        --image-id ami-e3106686 \
        --instance-type c4.large \
        --subnet-id ${SUBNET_ID} \
        --security-group-ids ${SGROUP_IDS} \
        --key-name ${KEY_NAME} \
        --associate-public-ip-address \
        --user-data "https://raw.githubusercontent.com/wmarinho/edw_cenipa/master/aws/user-data.sh" \
        --count 1
```

Obrigado!

Agradecimentos:
Marcelo Módolo – Globosat
Caio Moreno – IT4Biz
Fernando Maia – IT4Biz

Projetos:
https://github.com/wmarinho/edw_cenipa
https://github.com/wmarinho/docker-pentaho
https://hub.docker.com/r/wmarinho/pentaho/