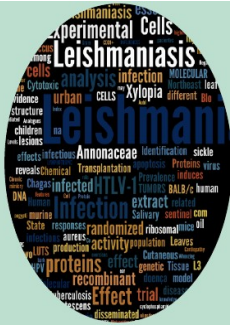
Big (and not so big) Data in Health Bahia'13



Infraestrutura computacional para suporte de aplicações de *big data* na área da Saúde

Maurício Barreto
Davide Rasella

Marcos Barreto

Centro Interdisciplinar em Ciências e Tecnologia da Informação

Infraestrutura computacional para suporte de aplicações de *big data* na área da Saúde

Objetivo

- Prover uma arquitetura de alto desempenho, robusta e adaptável para o processamento de aplicações intensivas de dados (*big data*).

Abordagem

- Explorar arquiteturas paralelas híbridas (multicore + manycore (GPU)) e/ou escaláveis.
- Prover serviços de dependabilidade e gestão autonômica da infraestrutura.
- Desenvolver portais de acesso para diferentes tipos de usuários com diferentes requisitos operacionais.

Infraestrutura computacional para suporte de aplicações de *big data* na área da Saúde

Equipe inicial

- CICTI / LaSiD
  - Prof. Marcos Barreto
  - Clicia Santos (mestranda)
  - Felipe Gutierrez (mestrando)
  - Robespierre Dantas (mestrando)
  - Pedro Novaes (bolsista IC)
  - Amanda Chagas (bolsista IC)
  - Marina Peixoto (voluntária)
  - Cristhian Carvalho (voluntário)

- Instituto de Saúde Coletiva (ISC)
  - Prof. Maurício Barreto
  - Prof. Davide Rasella

Contextualização

Estudo de caso (projeto piloto):

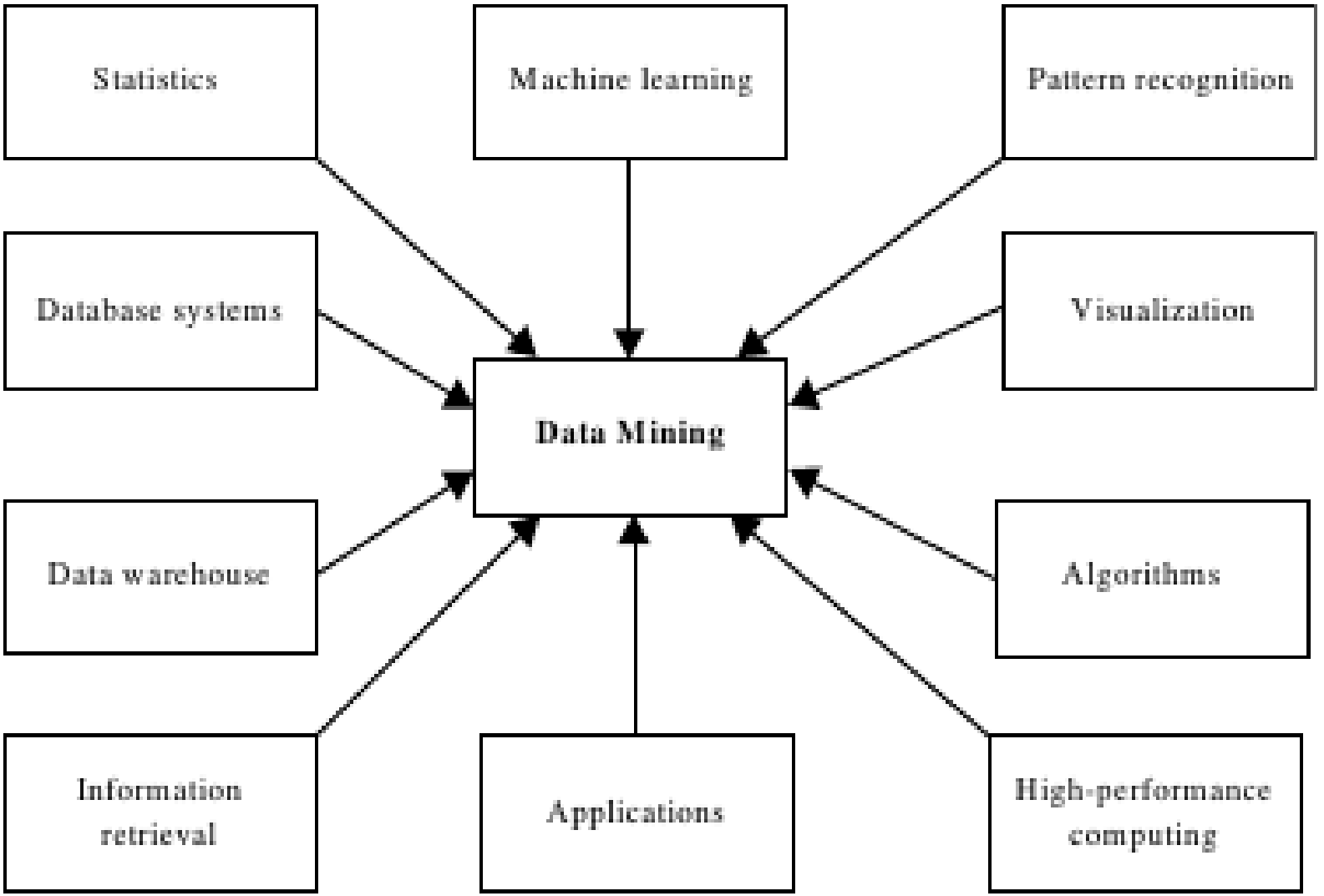
- Desenvolvimento de plataforma de estudos e avaliações permanentes dos efeitos do Bolsa Família e de outros Programas Sociais sobre a saúde, educação, trabalho e relações de gênero com base em coorte populacional referenciada no Cadastro Único.

Abordagem:

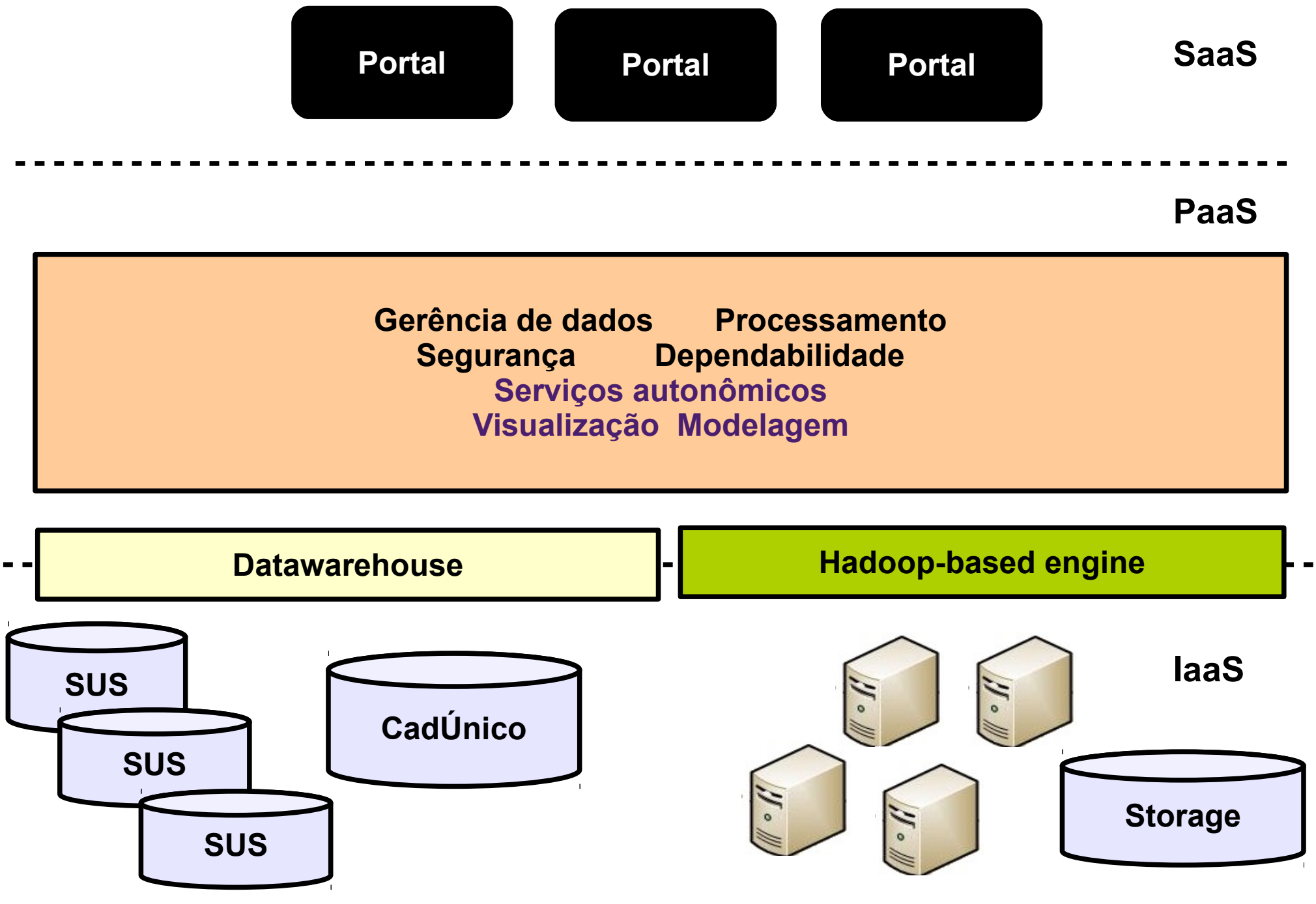
- Constituição de uma coorte populacional a partir do CadÚnico + bases SUS (SINAN, SIH e SIM).
- Processo progressivo de *linkage* de outras bases de dados.
- Testes com diferentes algoritmos probabilísticos de *linkage.*
- Abordagem MapReduce com base no Hadoop e outras ferramentas.
- Avaliação de métricas de acurácia, desempenho, consumo de recursos etc.
- Oferta de portais interativos para definição e monitoramento de aplicações.

Contextualização

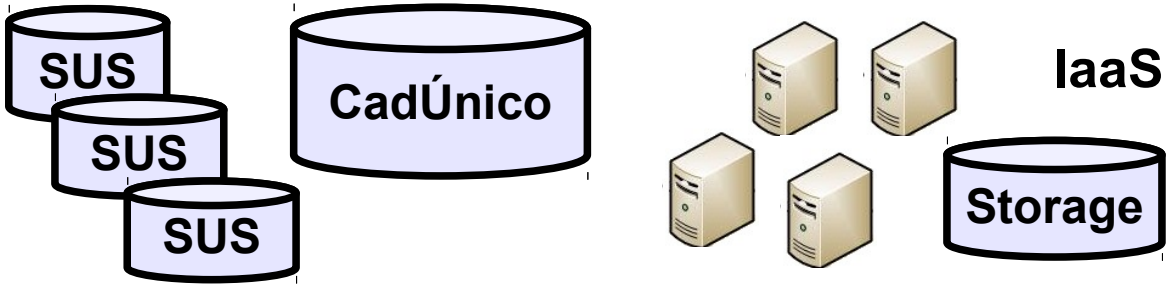
Algumas áreas envolvidas



Arquitetura proposta



Arquitetura proposta (2)



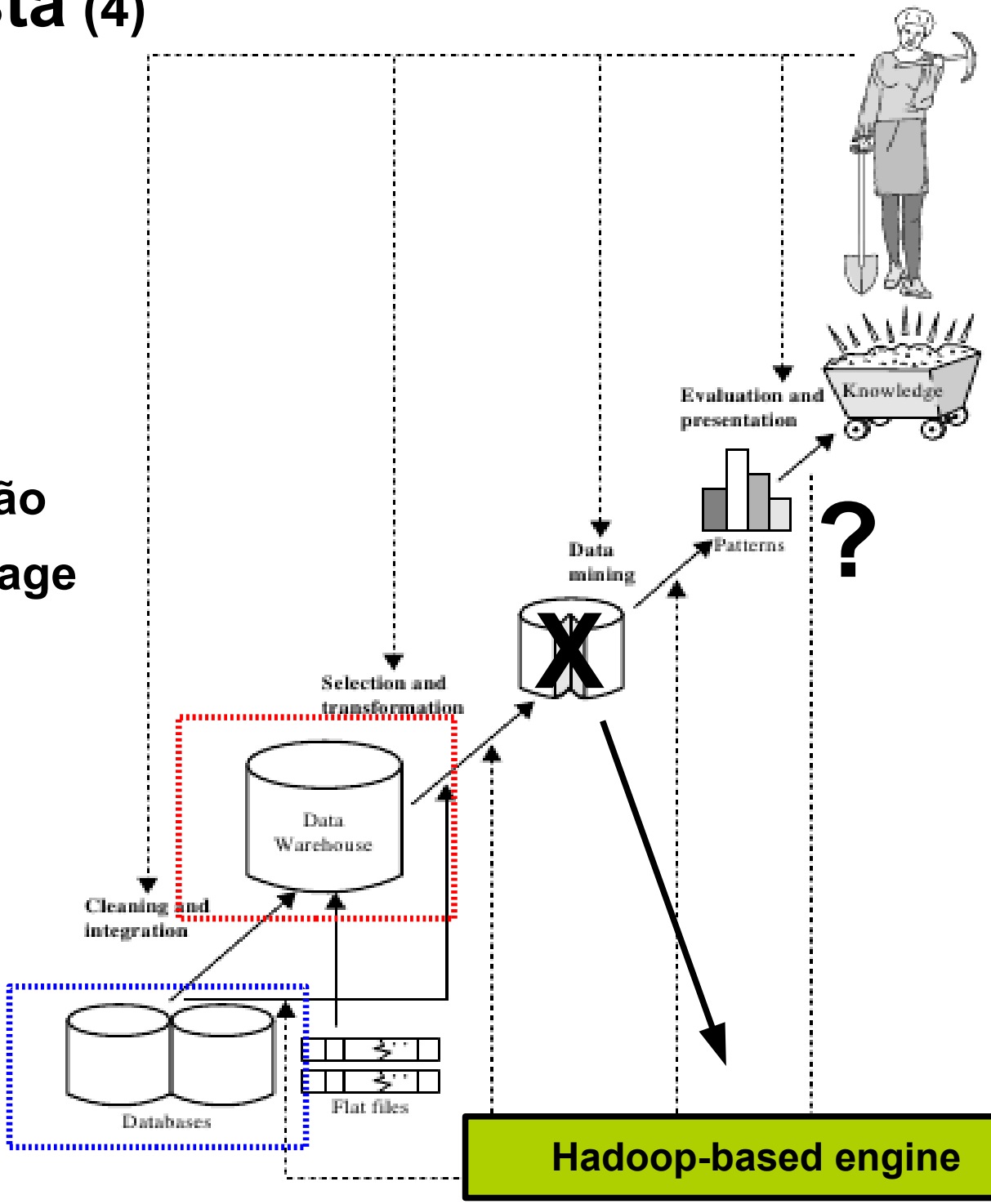
- **Nível IaaS**
  - Bases (*subset*) do SUS e CadÚnico
    - Transactional (ok) => analytical (?)
  - Clusters e servidores baseados em Hadoop
    - Modelo de execução + motor de inferência (algoritmos + rotinas analíticas)
  - Storage
    - Usuários, consultas (*reproducibility*), metadados, esquemas de visualização etc.
    - Armazenamento de resultados de linkage (data marts + algoritmos + parâmetros operacionais)
    - Download para recursos locais (análise estatística posterior)

Arquitetura proposta (3)

- **Nível PaaS - Datawarehouse**
  - Problemas
    - Falta de identificador único para integração das bases
    - Requisitos de confidencialidade
    - Crescimento incremental
    - Dados em formato TXT e/ou DBF

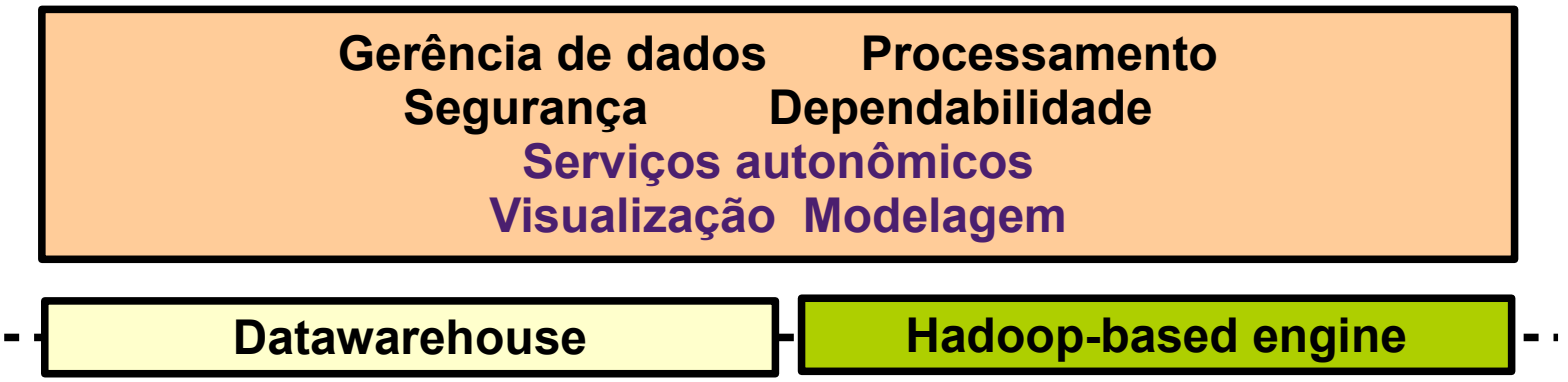
Arquitetura proposta (4)

- **Extract**
  - Carga de dados
- **Transform**
  - Anonimização
  - Avaliação e ponderação de parâmetros de linkage
- **Load**
  - Importação HDFS
  - => indefinição:
  - todas as etapas ou
  - somente a etapa
  - de linkage?



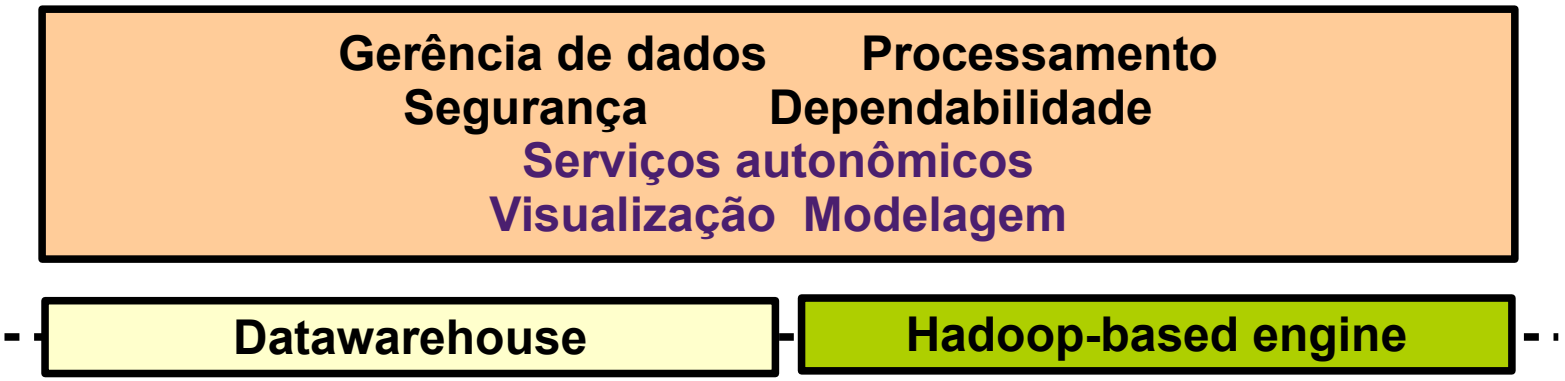
Arquitetura proposta (5)

- **Nível PaaS - Serviços**
  - Gerência de dados
    - Interação com camada de datawarehouse
    - Interação com *storage* para serviços de armazenamento



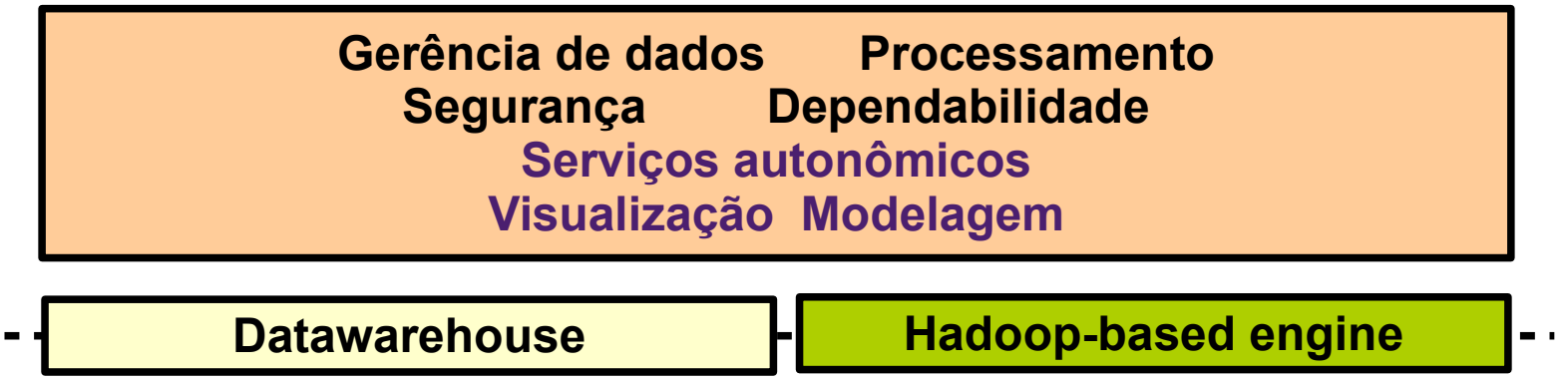
Arquitetura proposta (6)

- **Nível PaaS - Serviços**
  - Processamento
    - Motor de inferência (algoritmos probabilísticos)
    - Interação com Hadoop-based engine
    - Parametrização de algoritmos



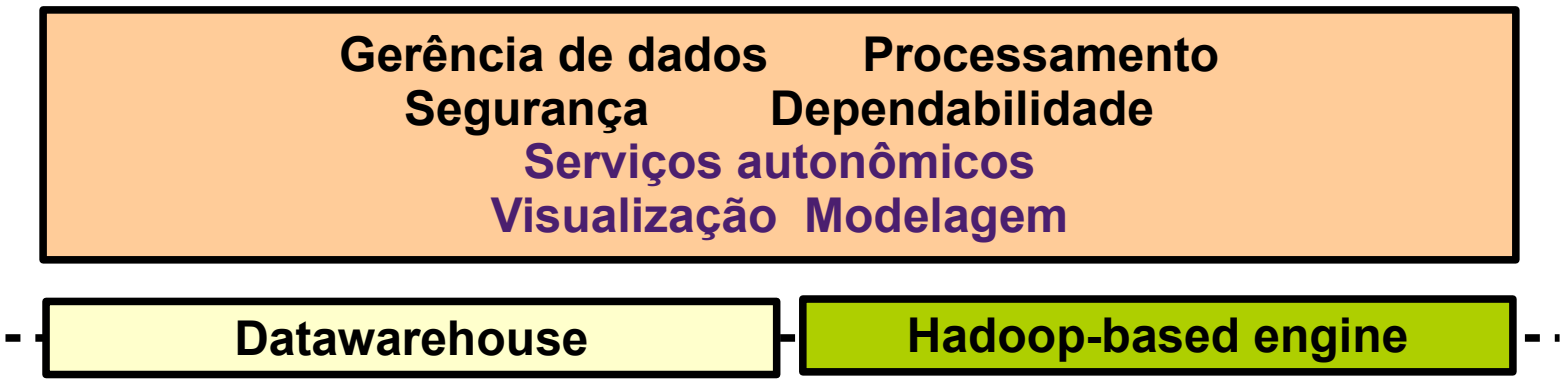
Arquitetura proposta (7)

- **Nível PaaS - Serviços**
  - Segurança
    - Autenticação de usuários
    - Conexão segura
    - Criptografia de dados (TrueCrypt)



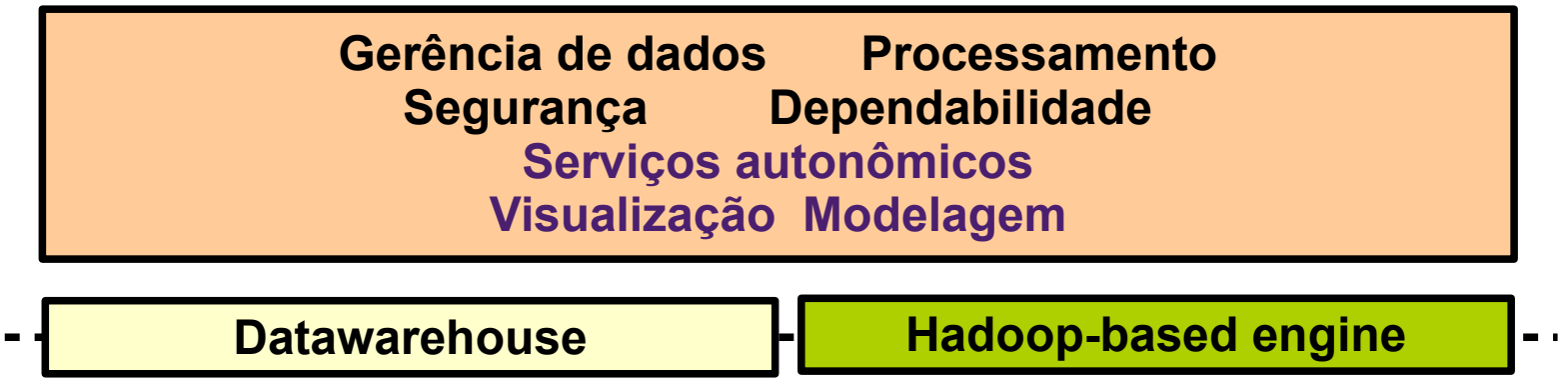
Arquitetura proposta (8)

- **Nível PaaS - Serviços**
  - Dependabilidade
    - Checkpoints no MapReduce?
    - Bases replicadas?



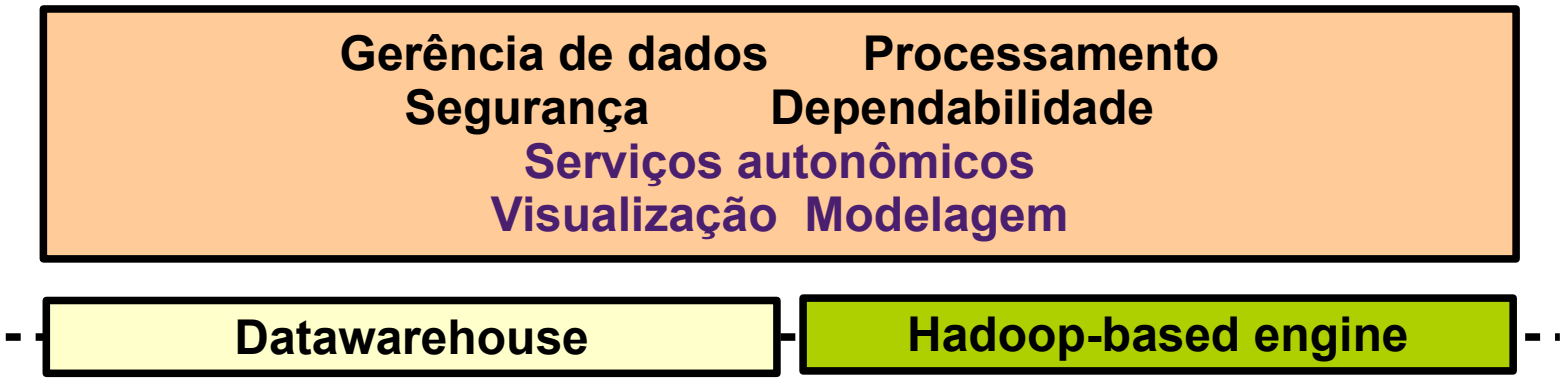
Arquitetura proposta (9)

- **Nível PaaS - Serviços**
  - Serviços autonômicos
    - Acordo de serviço (SLA) para cada portal?
    - Monitoramento de execução (bases de dados e VMs)
    - Just-in time resource provisioning
    - Auto-ajuste dos parâmetros de execução dos algoritmos => alterações no SLA?
    - Auto-ajuste nos algoritmos (motor de inferência)?



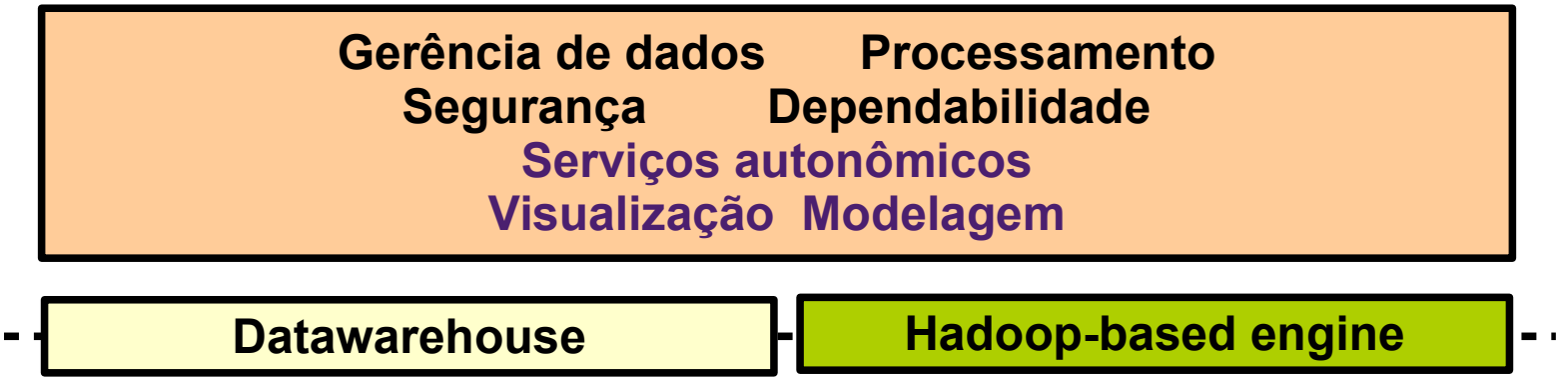
Arquitetura proposta (10)

- **Nível PaaS - Serviços**
  - Modelagem
    - Apoio para a construção de visões de dados (para cada portal).
    - Engenho de portais => esqueleto básico de portal customizável
    - Apoio para a visualização de dados
    - Metadados + ontologias



Arquitetura proposta (11)

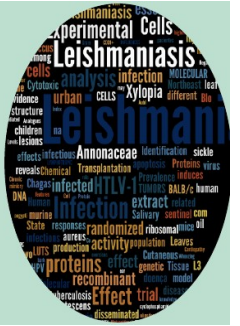
- **Nível PaaS - Serviços**
  - Visualização de dados
    - Modelos de visualização?
      - Escalar, volumes, tensorial, vetorial etc.
    - Interação/incorporação com ferramentas de visualização?
    - Geração de relatórios e gráficos etc.
    - Wiki com atividades e documentos do grupo (acesso aberto).



Arquitetura proposta (12)

- **Nível SaaS - Portais**
  - Construção de portais de acesso para cada área de interesse / usuários potenciais
    - Saúde
    - Educação
    - Trabalho ?
  - Acesso via Web
  - Aplicativo para dispositivo móvel?

Big (and not so big) Data in Health Bahia'13



Obrigado!

Contato:

marcoseb@dcc.ufba.br

LaSiD UFBA

Centro Interdisciplinar em Ciências e Tecnologia da Informação