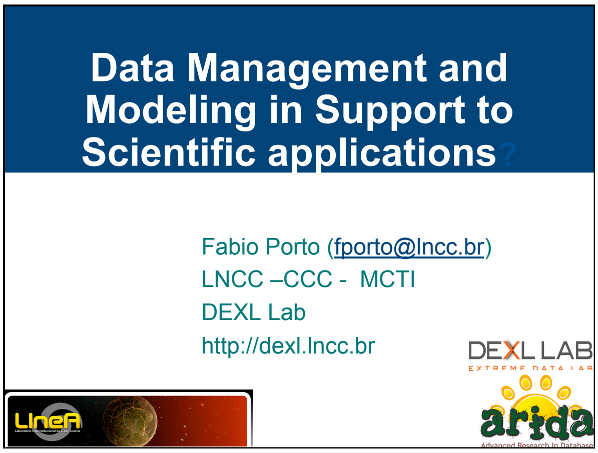


Data Management and Modeling in Support to Scientific applications?

Fabio Porto (fporto@lncc.br)
LNCC –CCC - MCTI
DEXL Lab
http://dexl.lncc.br

Outline

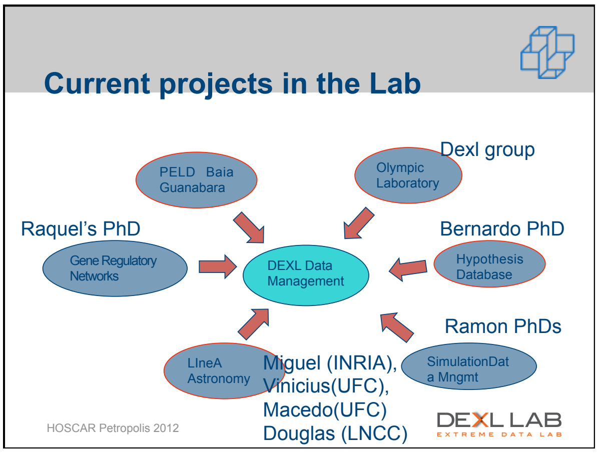
- Introduction
- Linked Science and Scientific Hypotheses
- Athletes Trajectory DW
- Astronomy integration as linked science
- QEF – Query Engine for data Intensive Applications
- Final remarks

HOSCAR Petropolis 2012

The DEXL Lab Mission

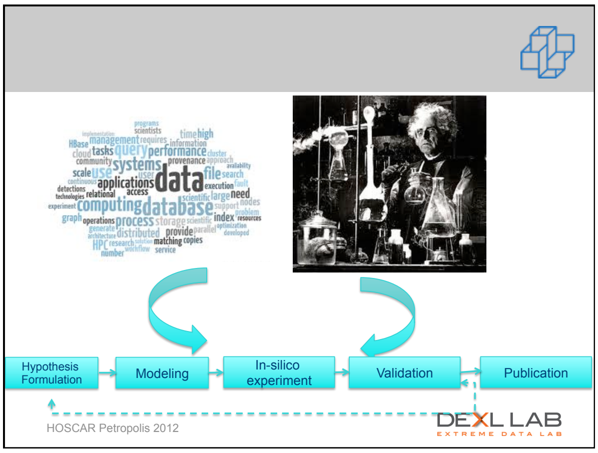
- To support in-silico science with data management techniques;
  - To develop interdisciplinary research with contributions on data modelling, design and management;
  - To develop tools and systems in support to in-silico science;

HOSCAR Petropolis 2012

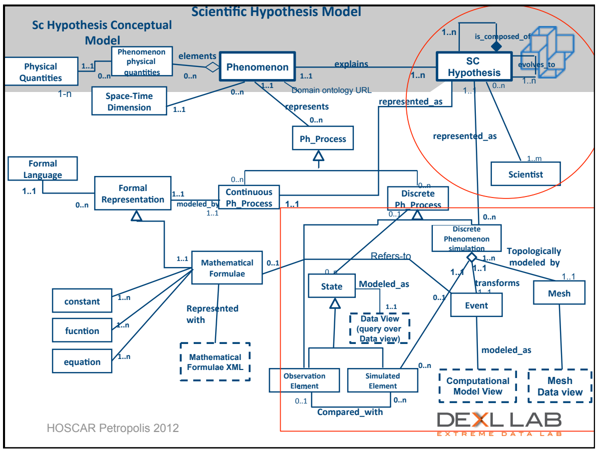


Current projects in the Lab

HOSCAR Petropolis 2012

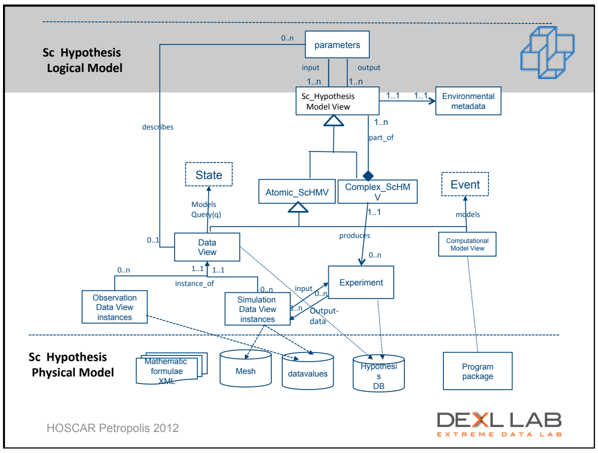


HOSCAR Petropolis 2012


Scientific Hypothesis Model

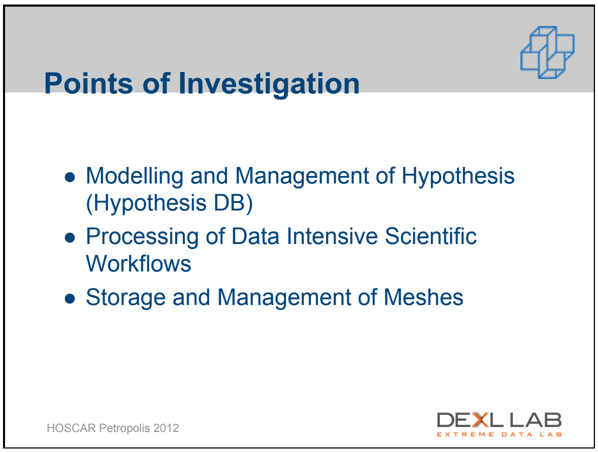
Sc Hypothesis Conceptual Model

HOSCAR Petropolis 2012


Sc Hypothesis Logical Model

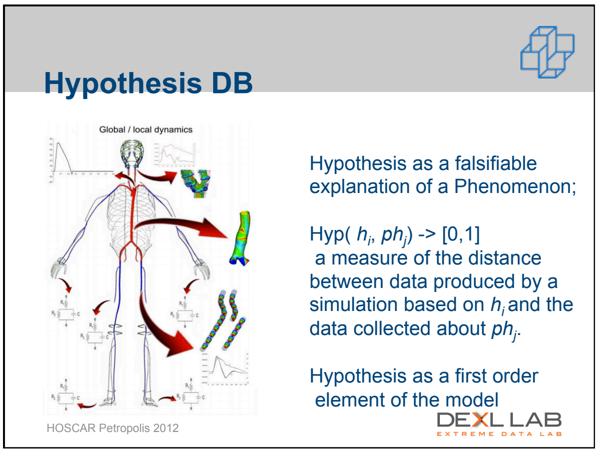
Sc Hypothesis Physical Model

HOSCAR Petropolis 2012


Points of Investigation

- Modelling and Management of Hypothesis (Hypothesis DB)
- Processing of Data Intensive Scientific Workflows
- Storage and Management of Meshes

HOSCAR Petropolis 2012



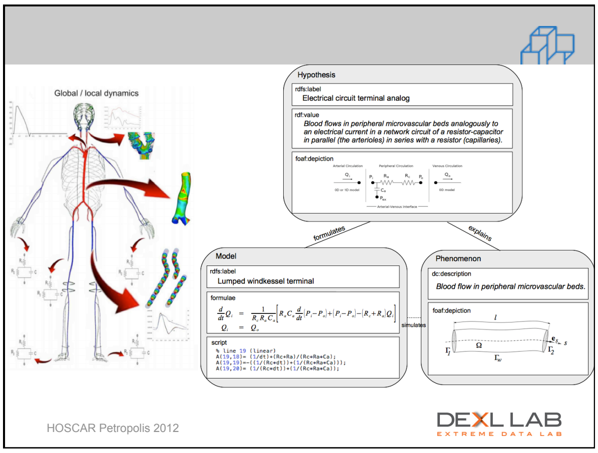
Hypothesis DB

Hypothesis as a falsifiable explanation of a Phenomenon;

Hyp(h_i, ph_j) -> [0,1]
a measure of the distance between data produced by a simulation based on h_i and the data collected about ph_j.

Hypothesis as a first order element of the model

HOSCAR Petropolis 2012



Hypothesis

rdfs:label
Electrical circuit terminal analog

rdf:value
Blood flows in peripheral microvascular beds analogously to an electrical current in a network circuit of a resistor-capacitor in parallel (the arterioles) in series with a resistor (capillaries).

foaf:depiction

formulates

explains

Model

rdfs:label
Lumped windkessel terminal

formulae

$$\frac{d}{dt}Q_i = \frac{1}{R_c R_a C_a}\left[R_a C_a \frac{d}{dt}(P_i - P_o) + (P_i - P_o) - (R_c + R_a)Q_i\right]$$
$$Q_i = Q_o$$

script

```
% line 19 (linear)
A(19,18)= (1/dt)+(Rc+Ra)/(Rc*Ra*Ca);
A(19,19)=-((1/(Rc*dt))+(1/(Rc*Ra*Ca)));
A(19,20)= (1/(Rc*dt))+(1/(Rc*Ra*Ca));
```

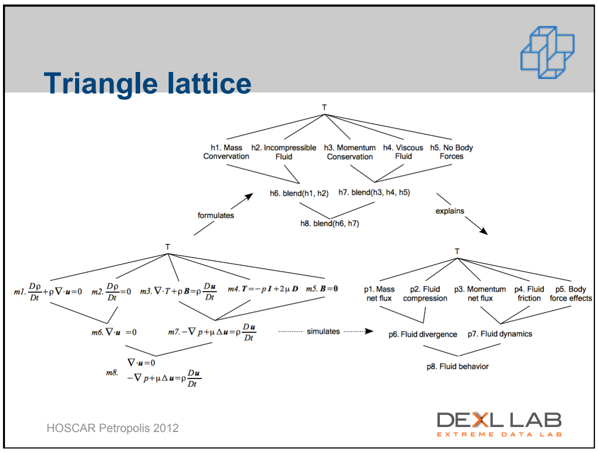
simulates

Phenomenon

dc:description
Blood flow in peripheral microvascular beds.

foaf:depiction

HOSCAR Petropolis 2012



Triangle lattice

T

h1. Mass Convervation

h2. Incompressible Fluid

h3. Momentum Conservation

h4. Viscous Fluid

h5. No Body Forces

h6. blend(h1, h2)

h7. blend(h3, h4, h5)

h8. blend(h6, h7)

formulates

explains

T

$m1.\ \frac{D\rho}{Dt} + \rho\nabla\cdot\boldsymbol{u} = 0$

$m2.\ \frac{D\rho}{Dt} = 0$

$m3.\ \nabla\cdot T + \rho\boldsymbol{B} = \rho\frac{D\boldsymbol{u}}{Dt}$

$m4.\ \boldsymbol{T} = -p\boldsymbol{I} + 2\mu\boldsymbol{D}$

$m5.\ \boldsymbol{B} = \boldsymbol{0}$

$m6.\ \nabla\cdot\boldsymbol{u} = 0$

$m7.\ -\nabla p + \mu\Delta\boldsymbol{u} = \rho\frac{D\boldsymbol{u}}{Dt}$

$m8.\ \nabla\cdot\boldsymbol{u} = 0$, $-\nabla p + \mu\Delta\boldsymbol{u} = \rho\frac{D\boldsymbol{u}}{Dt}$

simulates

T

p1. Mass net flux

p2. Fluid compression

p3. Momentum net flux

p4. Fluid friction

p5. Body force effects

p6. Fluid divergence

p7. Fluid dynamics

p8. Fluid behavior

HOSCAR Petropolis 2012

Linked Science

- An initiative to have a machine-readable content describing the scientific exploration;
- Support reproducibility of experiments;
- To foster reusing previous results;
- The community needs a more "open" science"

HOSCAR Petropolis 2012

Linked Science (or Linked Open Science)

- Is an initiative to interconnect all scientific assets;
- It is a combination of:
  - Linked data, semantic web
  - Open source;
  - Scientific workflows and provenance (OPM);
  - Scientific models;
  - Cloud computing;
  - …

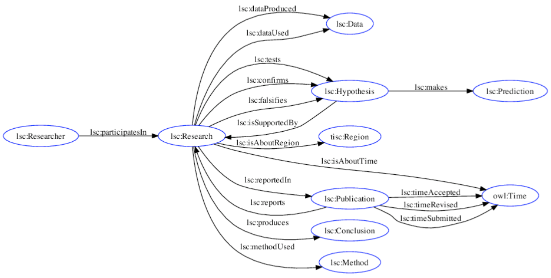
DEXL LAB
EXTREME DATA LAB

Linked Science Core Vocabulary (LSC)

- Defines a vocabulary (LSC) with “basic” terms for science;
  - More specific terminology shall be added by individual communities (minimal ontological commitment)

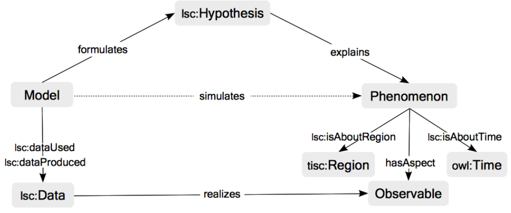
DEXL LAB
EXTREME DATA LAB

LSC Core Vocabulary

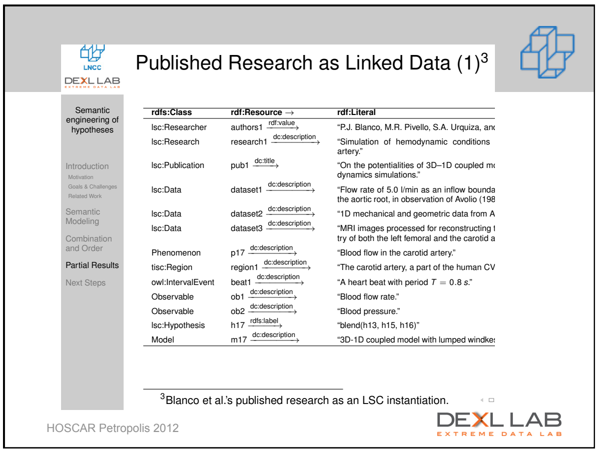


EXTREME DATA LAB

Extension to LSC



DEXL LAB
EXTREME DATA LAB



LNCC
DEXL LAB
EXTREME DATA LAB

Published Research as Linked Data (1)[3]

Semantic engineering of hypotheses

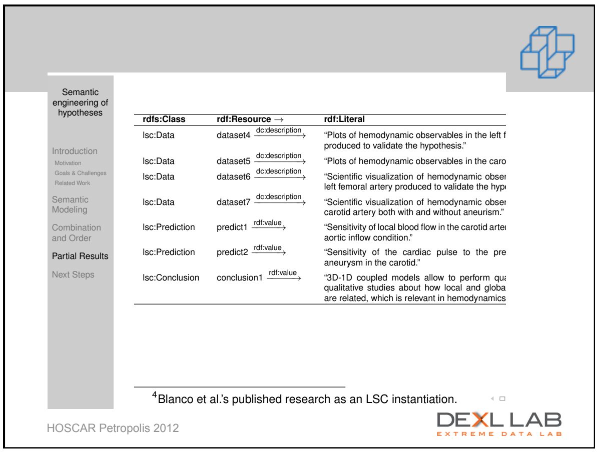
Introduction
Motivation
Goals & Challenges
Related Work
Semantic Modeling
Combination and Order
Partial Results
Next Steps

rdfs:Class	rdf:Resource →	rdf:Literal
lsc:Researcher	authors1 $\xrightarrow{\text{rdf:value}}$	"P.J. Blanco, M.R. Pivello, S.A. Urquiza, anc
lsc:Research	research1 $\xrightarrow{\text{dc:description}}$	"Simulation of hemodynamic conditions artery."
lsc:Publication	pub1 $\xrightarrow{\text{dc:title}}$	"On the potentialities of 3D–1D coupled mc dynamics simulations."
lsc:Data	dataset1 $\xrightarrow{\text{dc:description}}$	"Flow rate of 5.0 l/min as an inflow bounda the aortic root, in observation of Avolio (198
lsc:Data	dataset2 $\xrightarrow{\text{dc:description}}$	"1D mechanical and geometric data from A
lsc:Data	dataset3 $\xrightarrow{\text{dc:description}}$	"MRI images processed for reconstructing t try of both the left femoral and the carotid a
Phenomenon	p17 $\xrightarrow{\text{dc:description}}$	"Blood flow in the carotid artery."
tisc:Region	region1 $\xrightarrow{\text{dc:description}}$	"The carotid artery, a part of the human CV
owl:IntervalEvent	beat1 $\xrightarrow{\text{dc:description}}$	"A heart beat with period $T = 0.8\ s$."
Observable	ob1 $\xrightarrow{\text{dc:description}}$	"Blood flow rate."
Observable	ob2 $\xrightarrow{\text{dc:description}}$	"Blood pressure."
lsc:Hypothesis	h17 $\xrightarrow{\text{rdfs:label}}$	"blend(h13, h15, h16)"
Model	m17 $\xrightarrow{\text{dc:description}}$	"3D-1D coupled model with lumped windkes

[3] Blanco et al.'s published research as an LSC instantiation.

HOSCAR Petropolis 2012

DEXL LAB
EXTREME DATA LAB



Semantic engineering of hypotheses

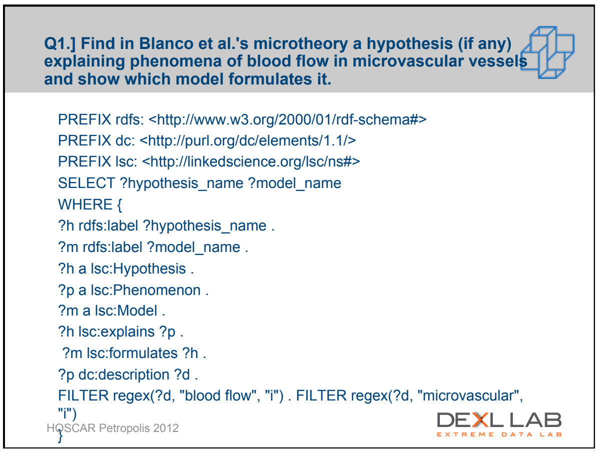
Introduction
Motivation
Goals & Challenges
Related Work
Semantic Modeling
Combination and Order
Partial Results
Next Steps

rdfs:Class	rdf:Resource →	rdf:Literal
lsc:Data	dataset4 $\xrightarrow{\text{dc:description}}$	"Plots of hemodynamic observables in the left f produced to validate the hypothesis."
lsc:Data	dataset5 $\xrightarrow{\text{dc:description}}$	"Plots of hemodynamic observables in the caro
lsc:Data	dataset6 $\xrightarrow{\text{dc:description}}$	"Scientific visualization of hemodynamic obser left femoral artery produced to validate the hyp
lsc:Data	dataset7 $\xrightarrow{\text{dc:description}}$	"Scientific visualization of hemodynamic obser carotid artery both with and without aneurism."
lsc:Prediction	predict1 $\xrightarrow{\text{rdf:value}}$	"Sensitivity of local blood flow in the carotid arte aortic inflow condition."
lsc:Prediction	predict2 $\xrightarrow{\text{rdf:value}}$	"Sensitivity of the cardiac pulse to the pre aneurysm in the carotid."
lsc:Conclusion	conclusion1 $\xrightarrow{\text{rdf:value}}$	"3D-1D coupled models allow to perform qua qualitative studies about how local and globa are related, which is relevant in hemodynamics

[4] Blanco et al.'s published research as an LSC instantiation.

HOSCAR Petropolis 2012

DEXL LAB
EXTREME DATA LAB

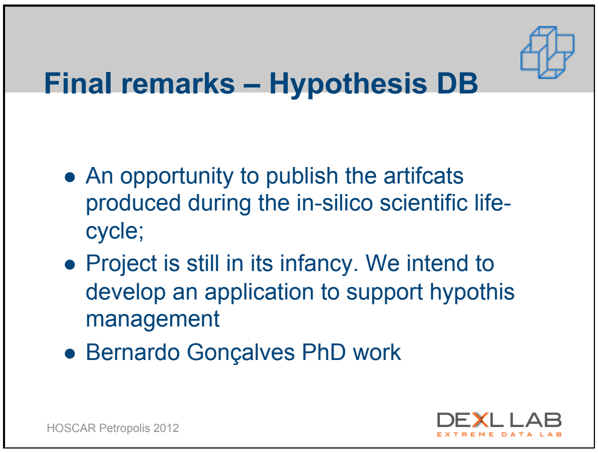


Q1.] Find in Blanco et al.'s microtheory a hypothesis (if any) explaining phenomena of blood flow in microvascular vessels and show which model formulates it.

```
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX dc: <http://purl.org/dc/elements/1.1/>
PREFIX lsc: <http://linkedscience.org/lsc/ns#>
SELECT ?hypothesis_name ?model_name
WHERE {
?h rdfs:label ?hypothesis_name .
?m rdfs:label ?model_name .
?h a lsc:Hypothesis .
?p a lsc:Phenomenon .
?m a lsc:Model .
?h lsc:explains ?p .
 ?m lsc:formulates ?h .
?p dc:description ?d .
FILTER regex(?d, "blood flow", "i") . FILTER regex(?d, "microvascular",
"i")
}
```

HOSCAR Petropolis 2012

DEXL LAB
EXTREME DATA LAB

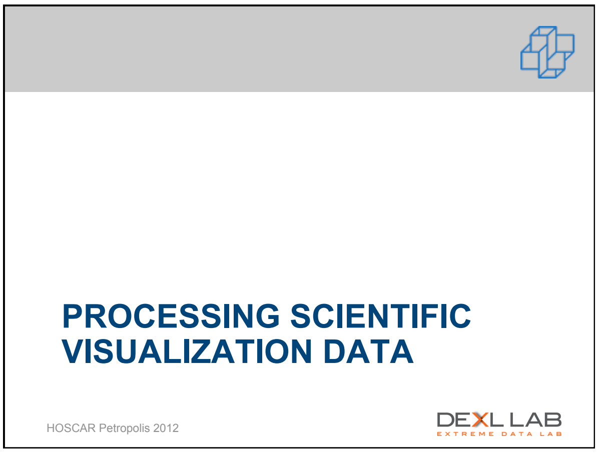


Final remarks – Hypothesis DB

- An opportunity to publish the artifcats produced during the in-silico scientific life-cycle;
- Project is still in its infancy. We intend to develop an application to support hypothis management
- Bernardo Gonçalves PhD work

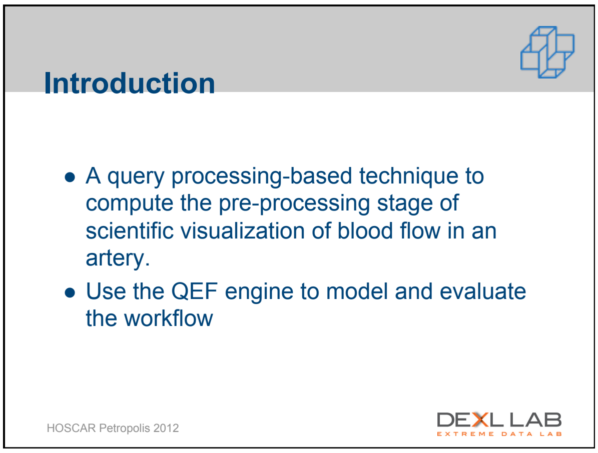
HOSCAR Petropolis 2012

DEXL LAB
EXTREME DATA LAB



PROCESSING SCIENTIFIC VISUALIZATION DATA

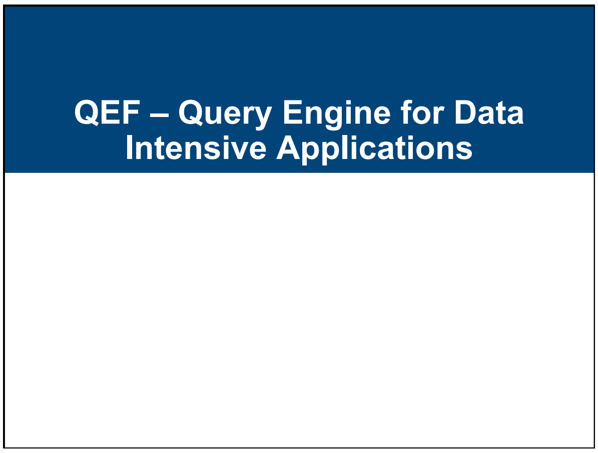
HOSCAR Petropolis 2012



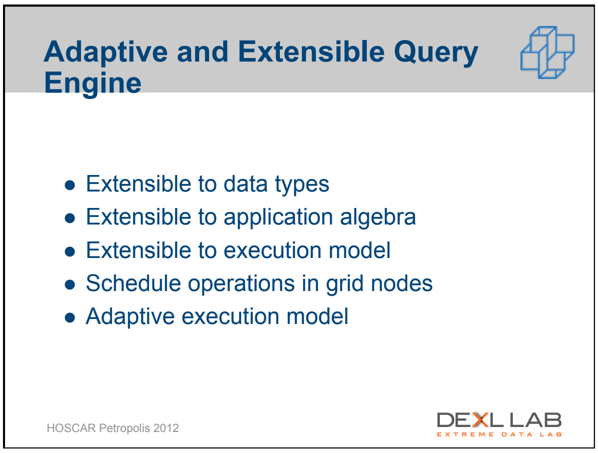
Introduction

- A query processing-based technique to compute the pre-processing stage of scientific visualization of blood flow in an artery.
- Use the QEF engine to model and evaluate the workflow

HOSCAR Petropolis 2012



QEF – Query Engine for Data Intensive Applications



Adaptive and Extensible Query Engine

- Extensible to data types
- Extensible to application algebra
- Extensible to execution model
- Schedule operations in grid nodes
- Adaptive execution model

HOSCAR Petropolis 2012

Objective

- Offer a query processing framework that can be extended to adapt to data intensive application needs;
- Offer transparency in using resources to answer queries;
  - Query optimization transparently introduced
  - Standardize remote communication using web services even when dealing with large amount of unstructured data
  - Run-time performance monitoring and decision

HOSCAR Petropolis 2012

DEXL LAB EXTREME DATA LAB

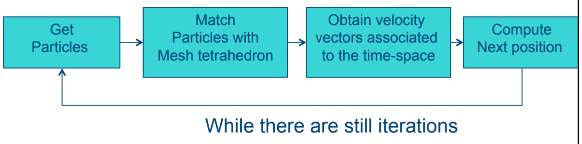
The problem

- Data sets
  - Mesh – tetrahedrons in 3D
  - Dataset of velocity and time
  - Virtual particles in an initial position
- Trajectories without collision
- A number of iterations through time-space

HOSCAR Petropolis 2012

DEXL LAB EXTREME DATA LAB

A scientific workflow


Get Particles
Match Particles with Mesh tetrahedron
Obtain velocity vectors associated to the time-space
Compute Next position
While there are still iterations

HOSCAR Petropolis 2012

DEXL LAB EXTREME DATA LAB

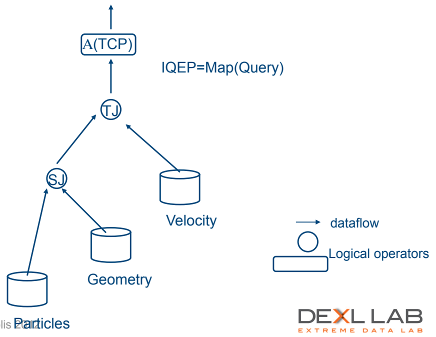
Modelling in QEF

- Each data set is a relation
  - Geometry (id, <3DPoint>)
  - Velocity (id, tetrahedronId,time,<velocity>)
  - Particle (id, iteratio, 3DPoint)
- Each activity of the workflow is an operator
  - Spatial-temporal join
  - Map (trajectory computing program)
- Add control operators
  - Orbit – to control iteration
  - Split/merge
  - Fold / unfold

HOSCAR Petropolis 2012

DEXL LAB EXTREME DATA LAB

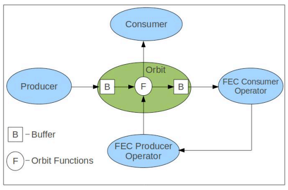
Initial logical QEP (IQEP)

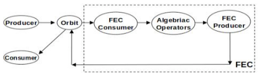


Control Operators

- Add data-flow and transformation operators
- Isolate application oriented operators from execution model data-flow concerns
- parallel grid based execution model:
  - **Split/Merge** - controls the routing of tuples to parallel nodes and the corresponding unification of multiple routes to a single flow
  - **Send/Receive** - marshalling/ unmarshalling of tuples and interface with communication mechanisms
  - **B2I/I2B** - blocks and unblocks tuples
  - **Orbit -** implements loop in a data-flow
  - **Fold/Unfold** - logical serialization of complex structues (e.g. PointList to Points)

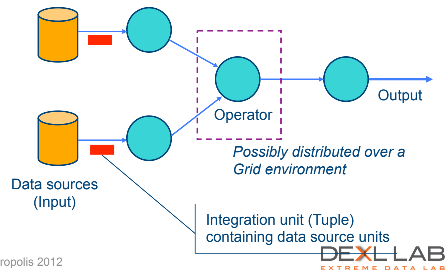
Orbit Model

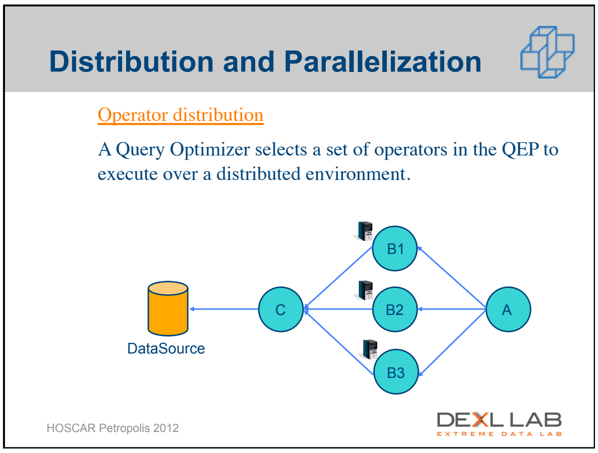




The Execution Model

Example of simple QEF Workflow

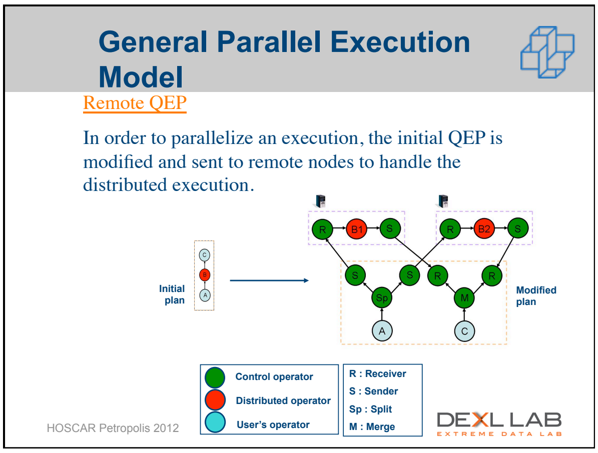

Operator
Output
Possibly distributed over a Grid environment
Data sources (Input)
Integration unit (Tuple) containing data source units



Distribution and Parallelization

Operator distribution

A Query Optimizer selects a set of operators in the QEP to execute over a distributed environment.

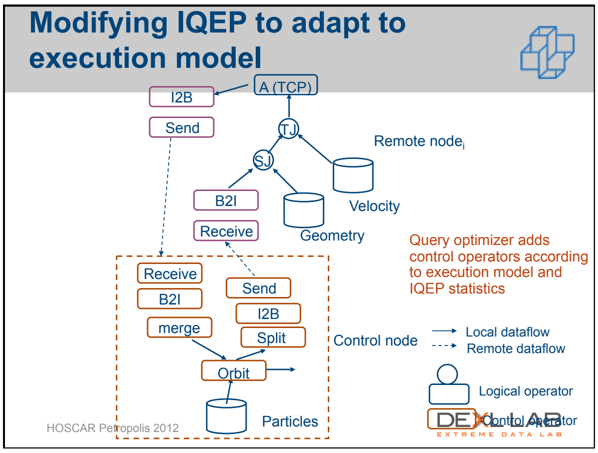


General Parallel Execution Model

Remote QEP

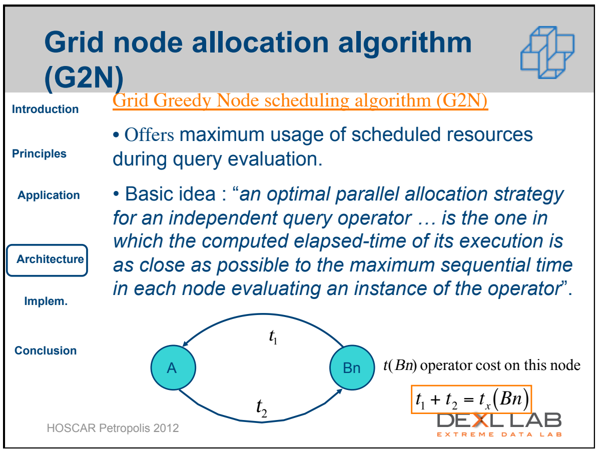
In order to parallelize an execution, the initial QEP is modified and sent to remote nodes to handle the distributed execution.

🟢 Control operator	R : Receiver
🔴 Distributed operator	S : Sender
🔵 User's operator	Sp : Split
	M : Merge



Modifying IQEP to adapt to execution model

Query optimizer adds control operators according to execution model and IQEP statistics



Grid node allocation algorithm (G2N)

Introduction

Principles

Application

Architecture

Implem.

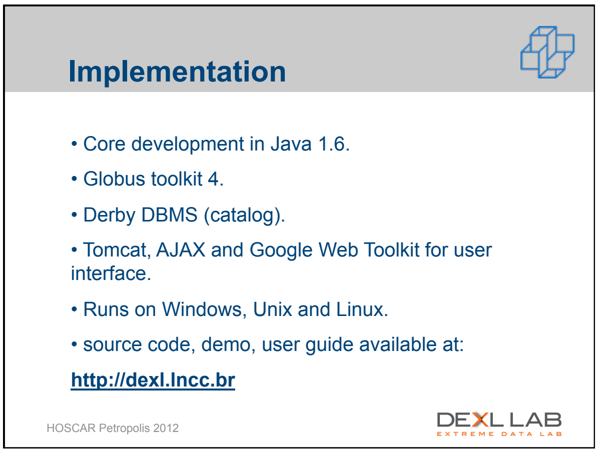
Conclusion

Grid Greedy Node scheduling algorithm (G2N)

- Offers maximum usage of scheduled resources during query evaluation.
- Basic idea : "*an optimal parallel allocation strategy for an independent query operator … is the one in which the computed elapsed-time of its execution is as close as possible to the maximum sequential time in each node evaluating an instance of the operator*".

$t(Bn)$ operator cost on this node

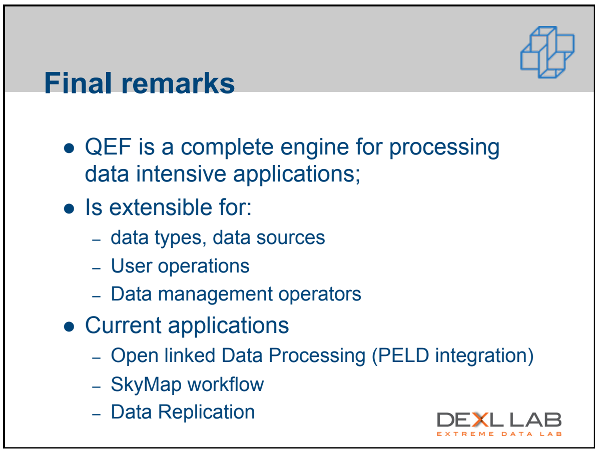
$$\boxed{t_1 + t_2 = t_x(Bn)}$$



Implementation

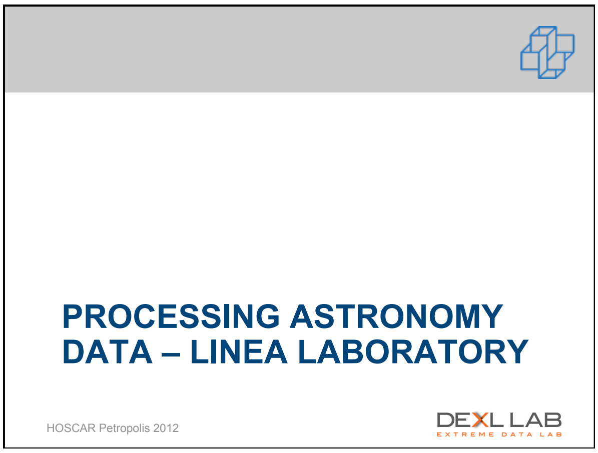
- Core development in Java 1.6.
- Globus toolkit 4.
- Derby DBMS (catalog).
- Tomcat, AJAX and Google Web Toolkit for user interface.
- Runs on Windows, Unix and Linux.
- source code, demo, user guide available at:

http://dexl.lncc.br

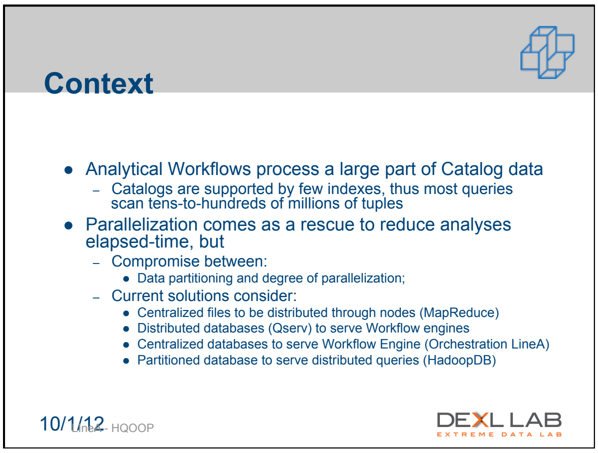


Final remarks

- QEF is a complete engine for processing data intensive applications;
- Is extensible for:
  - data types, data sources
  - User operations
  - Data management operators
- Current applications
  - Open linked Data Processing (PELD integration)
  - SkyMap workflow
  - Data Replication

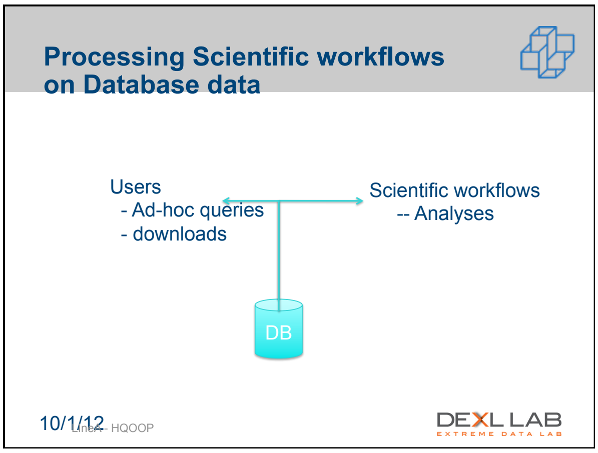


PROCESSING ASTRONOMY DATA – LINEA LABORATORY



Context

- Analytical Workflows process a large part of Catalog data
  - Catalogs are supported by few indexes, thus most queries scan tens-to-hundreds of millions of tuples
- Parallelization comes as a rescue to reduce analyses elapsed-time, but
  - Compromise between:
    - Data partitioning and degree of parallelization;
  - Current solutions consider:
    - Centralized files to be distributed through nodes (MapReduce)
    - Distributed databases (Qserv) to serve Workflow engines
    - Centralized databases to serve Workflow Engine (Orchestration LineA)
    - Partitioned database to serve distributed queries (HadoopDB)



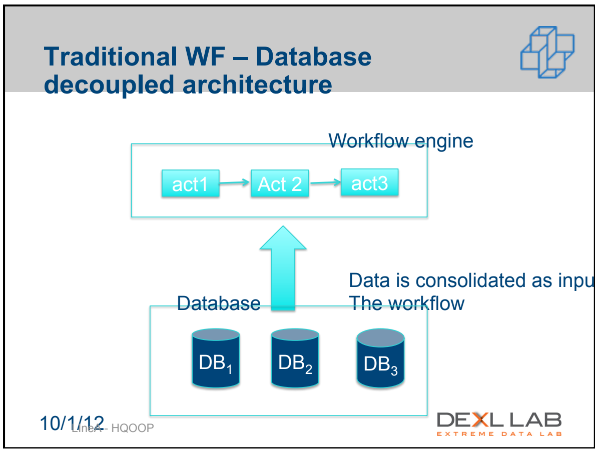
Processing Scientific workflows on Database data

Users
- Ad-hoc queries
- downloads

Scientific workflows
-- Analyses

DB

10/1/12 LIneA - HQOOP



Traditional WF – Database decoupled architecture

Workflow engine

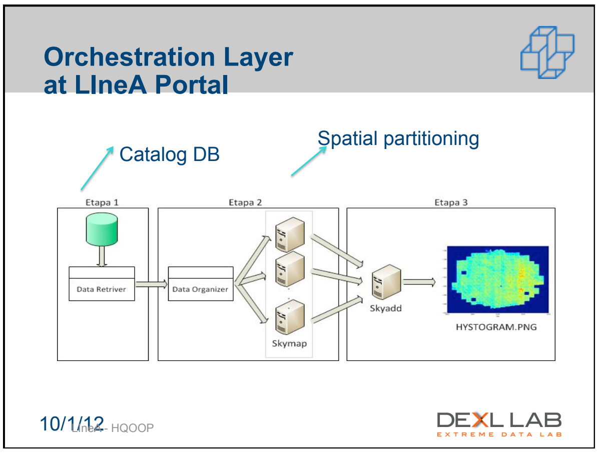
act1 → Act 2 → act3

Data is consolidated as inpu
The workflow

Database

DB_1 DB_2 DB_3

10/1/12 LIneA - HQOOP

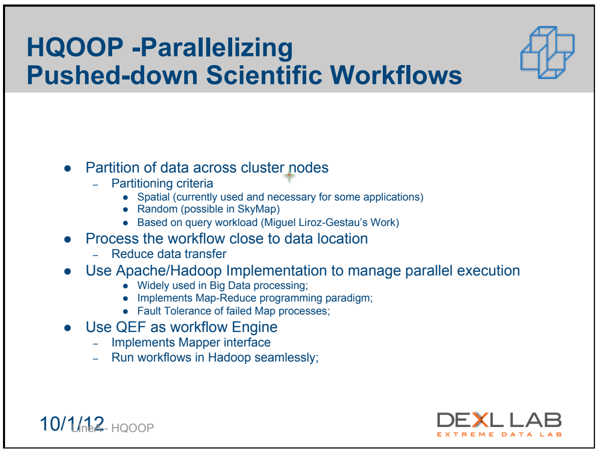


Orchestration Layer at LIneA Portal

Catalog DB

Spatial partitioning

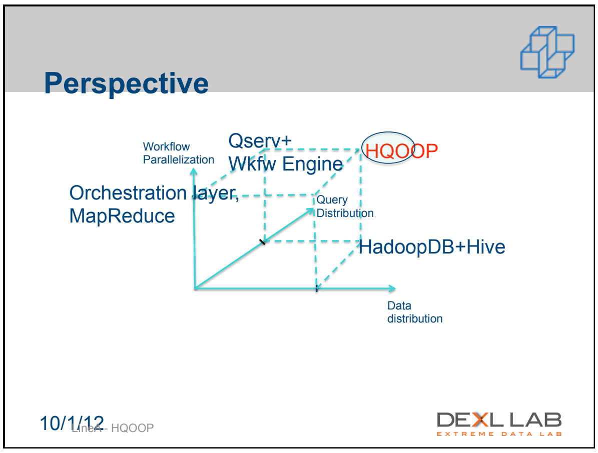
10/1/12 LIneA - HQOOP



HQOOP -Parallelizing Pushed-down Scientific Workflows

- Partition of data across cluster nodes
  - Partitioning criteria
    - Spatial (currently used and necessary for some applications)
    - Random (possible in SkyMap)
    - Based on query workload (Miguel Liroz-Gestau's Work)
- Process the workflow close to data location
  - Reduce data transfer
- Use Apache/Hadoop Implementation to manage parallel execution
    - Widely used in Big Data processing;
    - Implements Map-Reduce programming paradigm;
    - Fault Tolerance of failed Map processes;
- Use QEF as workflow Engine
  - Implements Mapper interface
  - Run workflows in Hadoop seamlessly;

10/1/12 LIneA - HQOOP



Perspective

Workflow Parallelization

Qserv+ Wkfw Engine

HQOOP

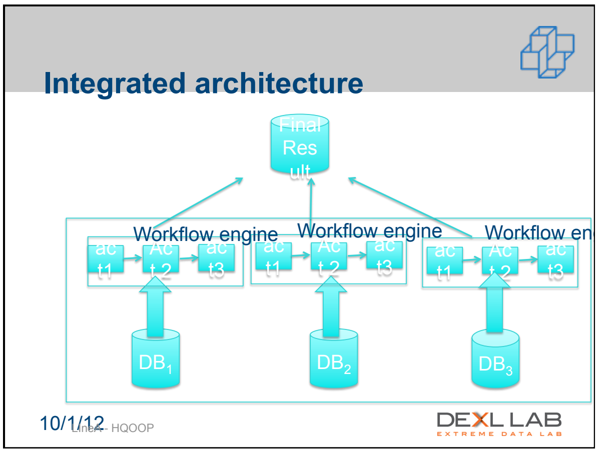
Orchestration layer, MapReduce

Query Distribution

HadoopDB+Hive

Data distribution

10/1/12 LineA - HQOOP



Integrated architecture

Final Res ult

Workflow engine

act1

Act 2

act3

Workflow engine

act1

Act 2

act3

Workflow en

act1

Act 2

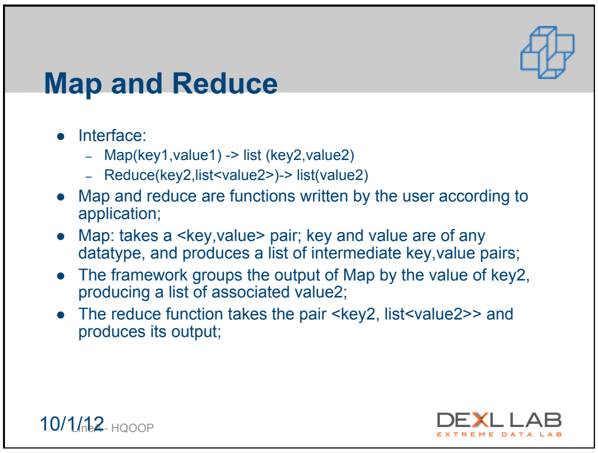
act3

DB_1

DB_2

DB_3

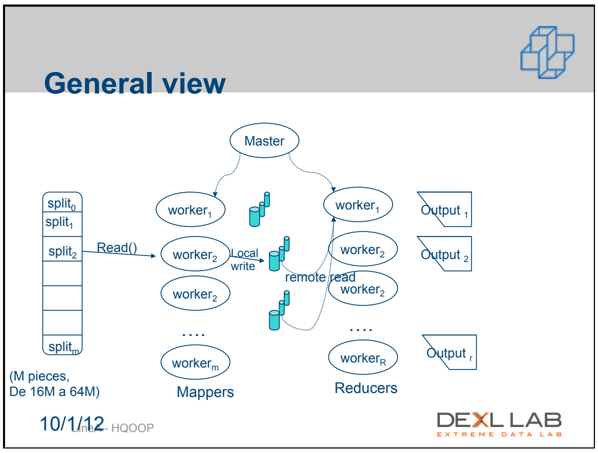
10/1/12 LineA - HQOOP



Map and Reduce

- Interface:
  - Map(key1,value1) -> list (key2,value2)
  - Reduce(key2,list<value2>)-> list(value2)
- Map and reduce are functions written by the user according to application;
- Map: takes a <key,value> pair; key and value are of any datatype, and produces a list of intermediate key,value pairs;
- The framework groups the output of Map by the value of key2, producing a list of associated value2;
- The reduce function takes the pair <key2, list<value2>> and produces its output;

10/1/12 LineA - HQOOP



General view

Master

$split_0$

$split_1$

$split_2$

$split_m$

Read()

$worker_1$

$worker_2$

$worker_2$

….

$worker_m$

Local write

remote read

$worker_1$

$worker_2$

$worker_2$

….

$worker_R$

Output 1

Output 2

Output r

(M pieces,
De 16M a 64M)

Mappers

Reducers

10/1/12 LineA - HQOOP

Process

- Initially, the MR framework splits the input into M partitions of fixed size;
- It initializes the Master node;
- The master node creates M+R workers and assign then maps and reduce functions, accordingly;
- Each Map reads its partition of the input and generates in memory its output;
- Periodically a process reads the buffer and groups the output values by key. It then writes the output to one of the R partitions, informing the master about its complete status and the partition addresses;
- Finally, the reduce reads each of its partitions and iterates over the keys, producing the results that are written to the output file.

Fault Tolerance

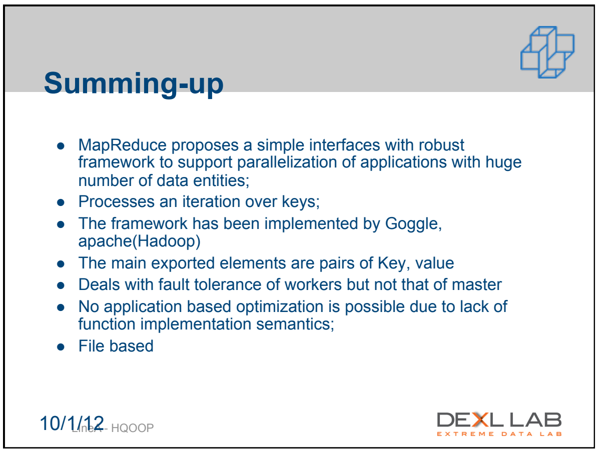
- Master keeps record of worker status (idle, in-progress, completed)
- It pings workers periodically
- It worker ping times-out, it is considered as dead and all completed work is re-scheduled to another node;

Task Granularity

- M and R much larger than the number of workers machines;
- Google defines M in terms of the size of the input partition (between 16M and 64M), and R a small multiple of the number of workers machine;
- Usual numbers:
  - Worker machines: 2000
  - Mappers: 200.000
  - Reducers: 5.000
- Reducers are in small number as they produce each an output file

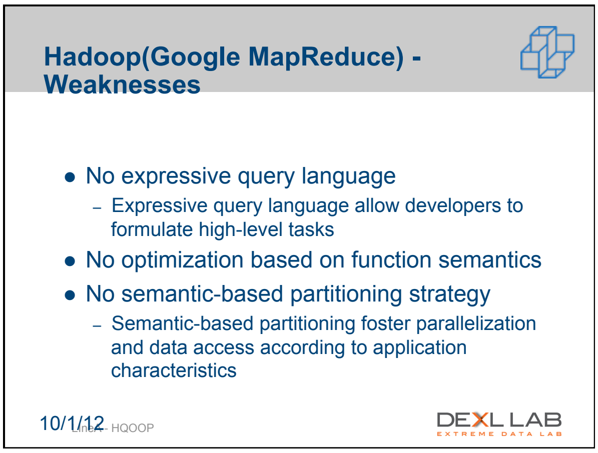
Partitioning of Intermediate results

- Intermediate results produced by map are re-partitioned into “R” fragments;
- Default partitioning function is:
  - Hash(key) mod R;
  - More semantically meaningful partition desired, if possible



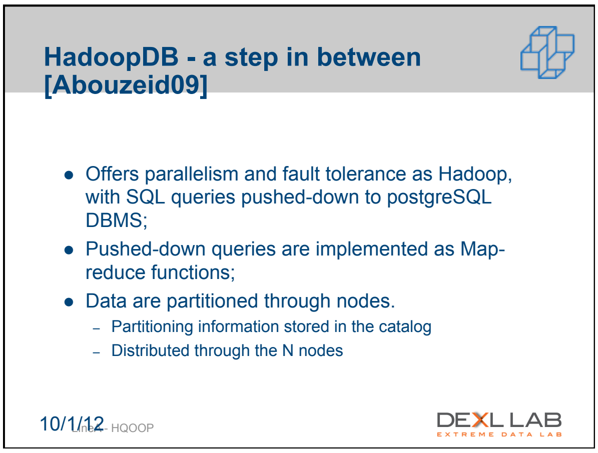
Summing-up

- MapReduce proposes a simple interfaces with robust framework to support parallelization of applications with huge number of data entities;
- Processes an iteration over keys;
- The framework has been implemented by Goggle, apache(Hadoop)
- The main exported elements are pairs of Key, value
- Deals with fault tolerance of workers but not that of master
- No application based optimization is possible due to lack of function implementation semantics;
- File based



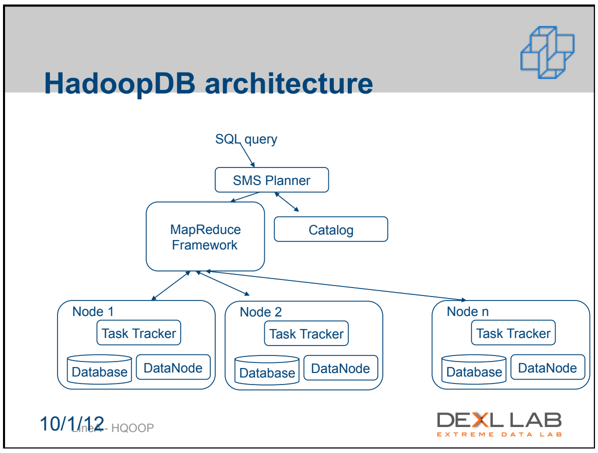
Hadoop(Google MapReduce) - Weaknesses

- No expressive query language
  - Expressive query language allow developers to formulate high-level tasks
- No optimization based on function semantics
- No semantic-based partitioning strategy
  - Semantic-based partitioning foster parallelization and data access according to application characteristics



HadoopDB - a step in between [Abouzeid09]

- Offers parallelism and fault tolerance as Hadoop, with SQL queries pushed-down to postgreSQL DBMS;
- Pushed-down queries are implemented as Map-reduce functions;
- Data are partitioned through nodes.
  - Partitioning information stored in the catalog
  - Distributed through the N nodes



HadoopDB architecture

SQL query

SMS Planner

MapReduce Framework

Catalog

Node 1

Task Tracker

Database

DataNode

Node 2

Task Tracker

Database

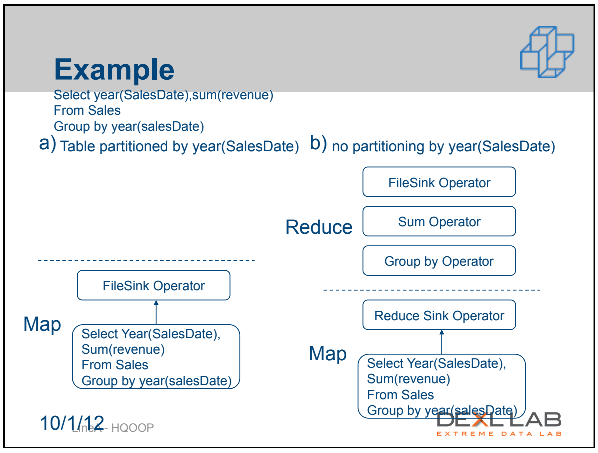
DataNode

Node n

Task Tracker

Database

DataNode



Example

Select year(SalesDate),sum(revenue)
From Sales
Group by year(salesDate)

a) Table partitioned by year(SalesDate)

Map

FileSink Operator

Select Year(SalesDate),
Sum(revenue)
From Sales
Group by year(salesDate)

b) no partitioning by year(SalesDate)

Reduce

FileSink Operator

Sum Operator

Group by Operator

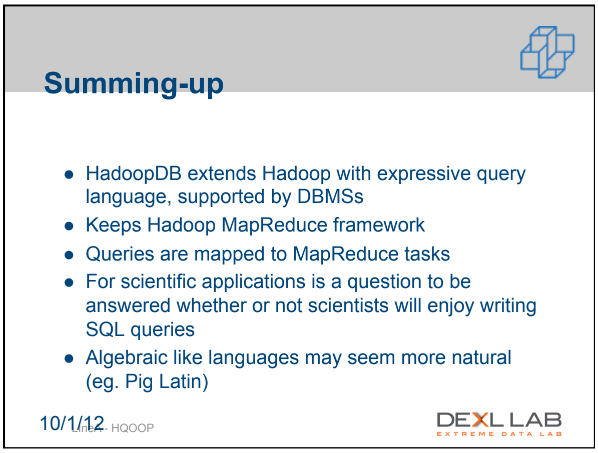
Map

Reduce Sink Operator

Select Year(SalesDate),
Sum(revenue)
From Sales
Group by year(salesDate)

10/1/12 LineA - HQOOP

DEXL LAB EXTREME DATA LAB

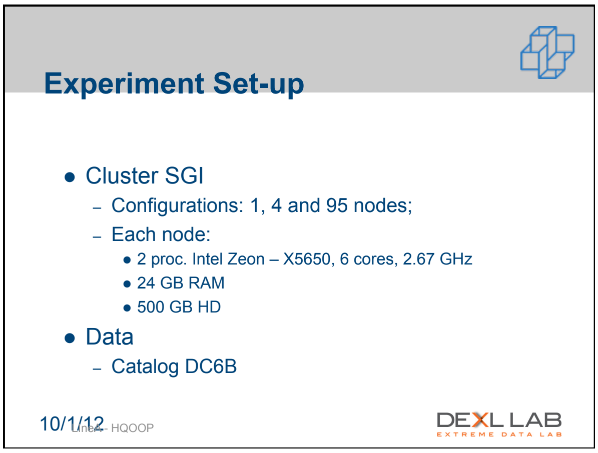


Summing-up

- HadoopDB extends Hadoop with expressive query language, supported by DBMSs
- Keeps Hadoop MapReduce framework
- Queries are mapped to MapReduce tasks
- For scientific applications is a question to be answered whether or not scientists will enjoy writing SQL queries
- Algebraic like languages may seem more natural (eg. Pig Latin)

10/1/12 LineA - HQOOP

DEXL LAB EXTREME DATA LAB

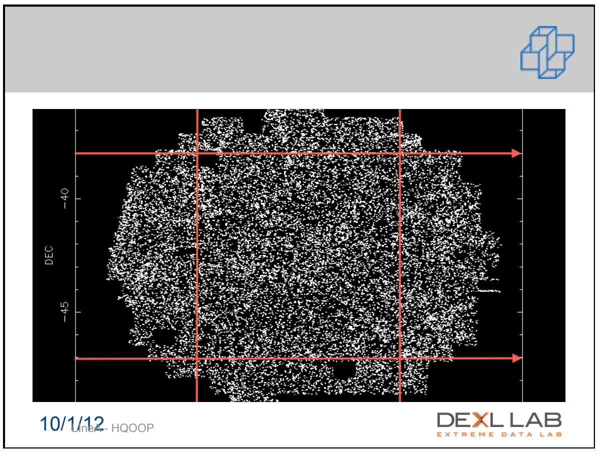


Experiment Set-up

- Cluster SGI
  - Configurations: 1, 4 and 95 nodes;
  - Each node:
    - 2 proc. Intel Zeon – X5650, 6 cores, 2.67 GHz
    - 24 GB RAM
    - 500 GB HD
- Data
  - Catalog DC6B

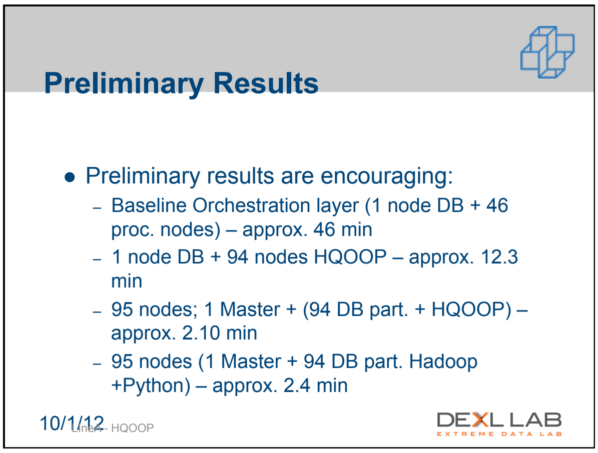
10/1/12 LineA - HQOOP

DEXL LAB EXTREME DATA LAB



10/1/12 LineA - HQOOP

DEXL LAB EXTREME DATA LAB

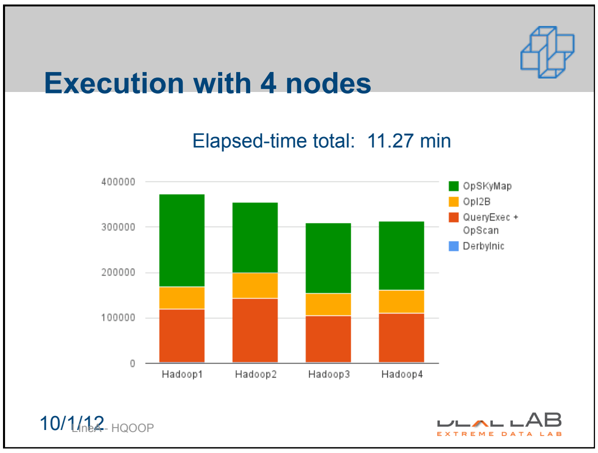


Preliminary Results

- Preliminary results are encouraging:
  - Baseline Orchestration layer (1 node DB + 46 proc. nodes) – approx. 46 min
  - 1 node DB + 94 nodes HQOOP – approx. 12.3 min
  - 95 nodes; 1 Master + (94 DB part. + HQOOP) – approx. 2.10 min
  - 95 nodes (1 Master + 94 DB part. Hadoop +Python) – approx. 2.4 min

10/1/12 LineA - HQOOP

DEXL LAB
EXTREME DATA LAB

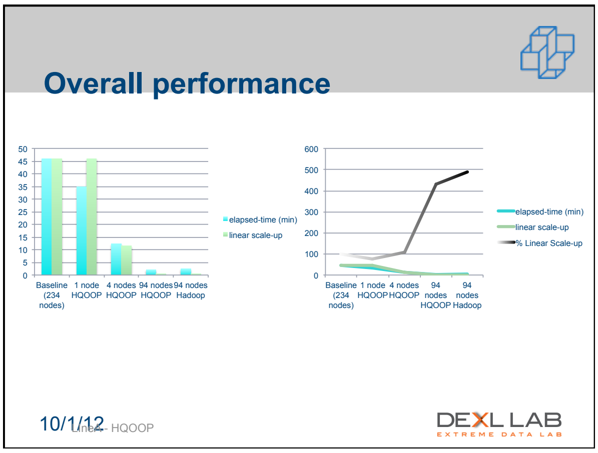


Execution with 4 nodes

Elapsed-time total: 11.27 min

10/1/12 LineA - HQOOP

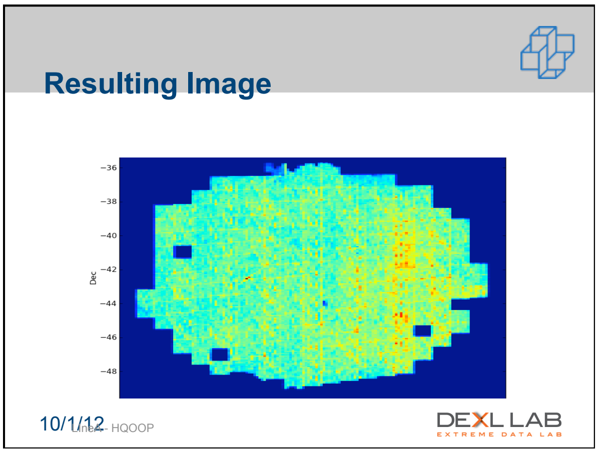
DEXL LAB
EXTREME DATA LAB



Overall performance

10/1/12 LineA - HQOOP

DEXL LAB
EXTREME DATA LAB



Resulting Image

10/1/12 LineA - HQOOP

DEXL LAB
EXTREME DATA LAB

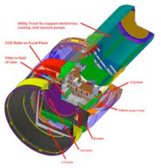
TLPP Processing

- Transfer Data
  - From telescope to data storage sites
- Load
  - Data Ingestion Procedures
  - Data Management
  - Data Replication
  - Data Model
  - Query Processing
- Process
  - Scientific Workflows
  - ★ Data locality
  - Provenance
  - Store of workflow results in DB
- Publish
  - Linked-Data

Final Remarks

- HQOOP shows interesting initial results
  - Evaluate with other pipelines
  - Enhancing data partitioning and load procedure
  - Allow workflow to be passed as parameter
  - Deal with fault tolerance
  - Evaluate other possible configurations

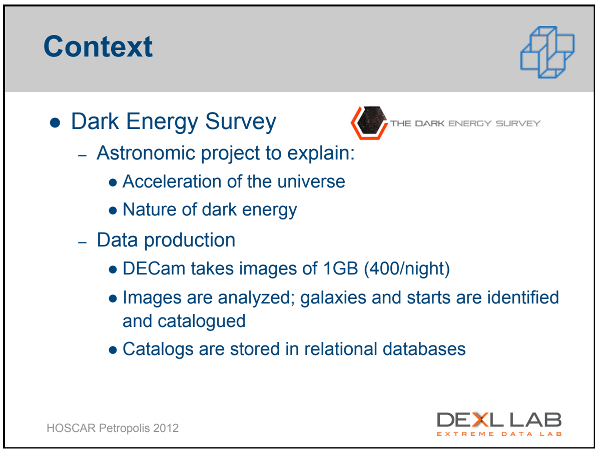
Dark Energy Survey and LSST – Large Synoptic Survey Telescope



Cerro Pachón – Future site of the LSST
LSST Rendering on El Peñón
SOAR
Gemini
Cerro Pachón ridge – view from northwest

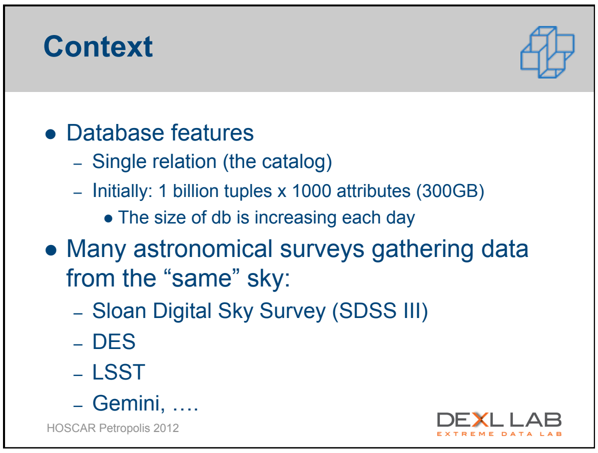
- 800 images p/ night during 10 anos !!
- Map 3D of the Universe
- 30 TeraBytes per night
- **30 PetaBytes in 10 years**

LSST – simulated image from sky



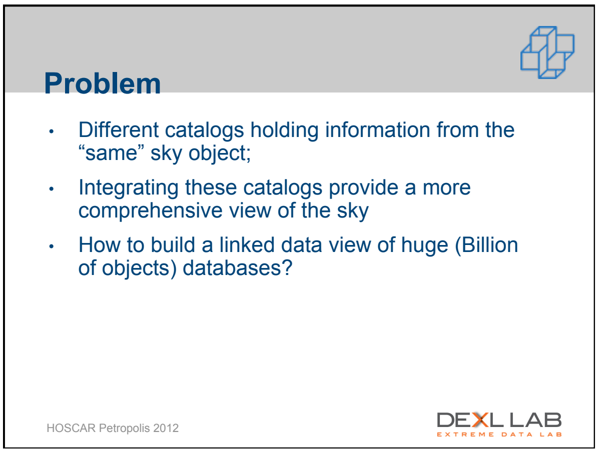
Context

- Dark Energy Survey
  - Astronomic project to explain:
    - Acceleration of the universe
    - Nature of dark energy
  - Data production
    - DECam takes images of 1GB (400/night)
    - Images are analyzed; galaxies and starts are identified and catalogued
    - Catalogs are stored in relational databases



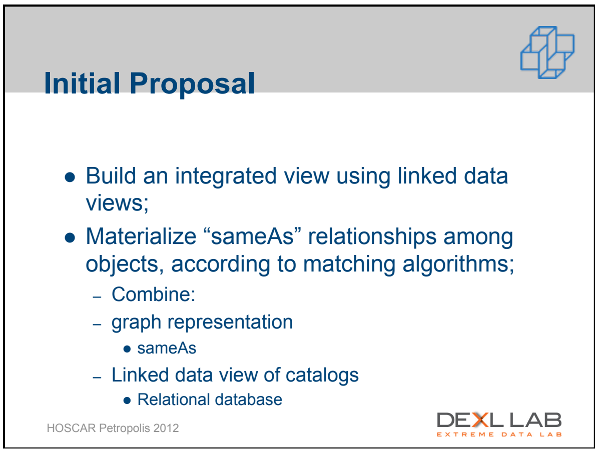
Context

- Database features
  - Single relation (the catalog)
  - Initially: 1 billion tuples x 1000 attributes (300GB)
    - The size of db is increasing each day
- Many astronomical surveys gathering data from the “same” sky:
  - Sloan Digital Sky Survey (SDSS III)
  - DES
  - LSST
  - Gemini, ….



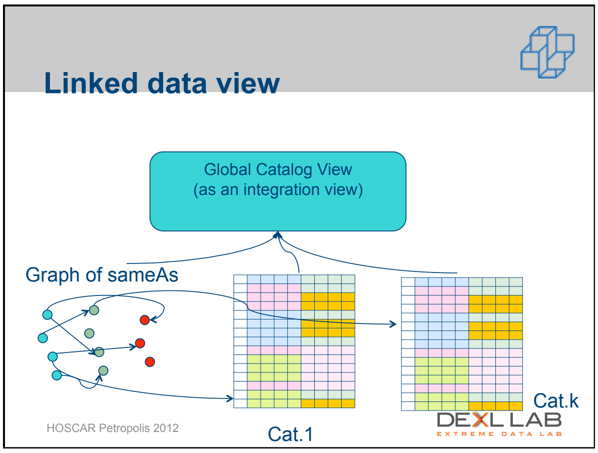
Problem

- Different catalogs holding information from the “same” sky object;
- Integrating these catalogs provide a more comprehensive view of the sky
- How to build a linked data view of huge (Billion of objects) databases?



Initial Proposal

- Build an integrated view using linked data views;
- Materialize “sameAs” relationships among objects, according to matching algorithms;
  - Combine:
  - graph representation
    - sameAs
  - Linked data view of catalogs
    - Relational database



Linked data view

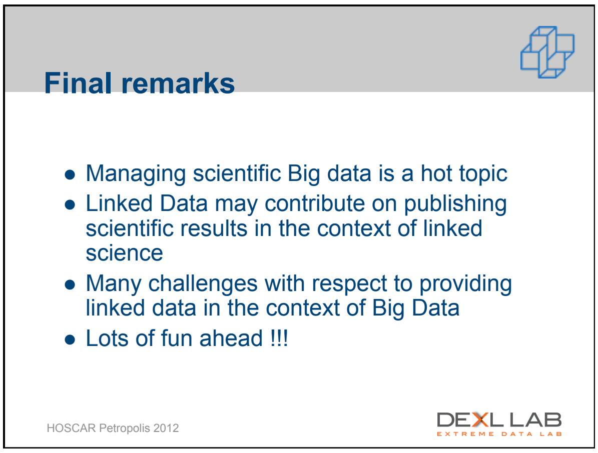
Global Catalog View
(as an integration view)

Graph of sameAs

Cat.1

Cat.k

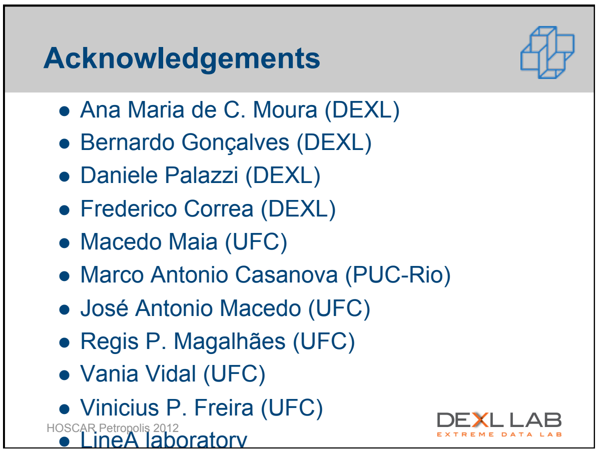
HOSCAR Petropolis 2012



Final remarks

- Managing scientific Big data is a hot topic
- Linked Data may contribute on publishing scientific results in the context of linked science
- Many challenges with respect to providing linked data in the context of Big Data
- Lots of fun ahead !!!

HOSCAR Petropolis 2012



Acknowledgements

- Ana Maria de C. Moura (DEXL)
- Bernardo Gonçalves (DEXL)
- Daniele Palazzi (DEXL)
- Frederico Correa (DEXL)
- Macedo Maia (UFC)
- Marco Antonio Casanova (PUC-Rio)
- José Antonio Macedo (UFC)
- Regis P. Magalhães (UFC)
- Vania Vidal (UFC)
- Vinicius P. Freira (UFC)
- LineA laboratory

HOSCAR Petropolis 2012

Obrigado !

HOSCAR Petropolis 2012