



Fundação Vanzolini

Dominando Big Data com o uso de Plataformas Gratuitas

Aula 6

Agenda da aula 6

- ✓ Resolução Chicago Crime
- ✓ Carregamento de dados
- ✓ Próximos passos

Desafio: Chicago Crimes

Desafio Chicago Crimes:

- Normalize o dataset de crimes em duas tabelas:
 - uma tabela de crimes e;
 - uma tabela de endereços (block, community area, district)
- Mantenha um campo de ligação entre as duas tabelas

row_id	day	time	case_number	iucr	primary_type	description	block_id
1	20150905	133000	HY411648	0486	BATTERY	DOMESTIC BATTERY SIMPLE	1
2	20150904	113000	HY411615	0870	THEFT	POCKET-PICKING	2
3	20180901	100	JC213529	0810	THEFT	OVER \$500	3
4	20150905	124500	HY411595	2023	NARCOTICS	POSS: HEROIN (BRN/TAN)	4
5	20150905	130000	HY411610	0560	ASSAULT	SIMPLE	5
6	20150905	105500	HY411435	0610	BURGLARY	FORCIBLE ENTRY	6
7	20150904	180000	HY411629	0620	BURGLARY	UNLAWFUL ENTRY	7
8	20150905	130000	HY411605	0860	THEFT	RETAIL THEFT	8
9	20150905	113000	HY411654	0320	ROBBERY	STRONGARM - NO WEAPON	9
10	20160501	2500	JC212333	1153	DECEPTIVE PRACTICE	FINANCIAL IDENTITY THEFT OVER \$ 300	10

block_id	block
1	043XX S WOOD ST
2	008XX N CENTRAL AVE
3	082XX S INGLESIDE AVE
4	035XX W BARRY AVE
5	0000X N LARAMIE AVE
6	082XX S LOOMIS BLVD
7	021XX W CHURCHILL ST
8	025XX W CERMAK RD
9	031XX W WASHINGTON BLVD
10	055XX S ROCKWELL ST

Solução proposta (BWR_RollupBlock.ecl):

```
IMPORT $;

Layout_T_recs := RECORD
    UNSIGNED4 block_ID := $.Formatted_File.row_id;
    $.Formatted_File.Block;
END;

T_recs := TABLE($.Formatted_File,Layout_T_recs);
S_recs := SORT(T_recs,block);
S_recs;
COUNT(S_recs);

Layout_T_recs RollCSV(Layout_T_recs L, Layout_T_recs R) := TRANSFORM
    SELF.block_ID := IF(L.block_ID < R.block_ID,L.block_ID,R.block_ID);
    SELF := L;
END;

Rollup_block := ROLLUP(S_Recs,LEFT.block=RIGHT.block,RollCSV(LEFT,RIGHT));

Rollup_block;
COUNT(Rollup_block);

S_Rollup_block := SORT(Rollup_block,block_ID);
OUTPUT(S_Rollup_block, '~CLASS::HMW::OUT::LookupBlock',OVERWRITE);
```

Solução proposta (NormAddrRecs.ecl):

```
EXPORT NormAddrRecs := MODULE

  EXPORT Layout := RECORD
    UNSIGNED4 block_ID;
    STRING38 Block;
  END;

  EXPORT File := DATASET('~CLASS::HMW::OUT::LookupBlock', Layout, THOR);
END;
```

Solução proposta (NormCrimeRecs.ecl):

```
IMPORT $;

EXPORT NormCrimeRecs := MODULE
  EXPORT Layout := RECORD
    UNSIGNED row_id;
    UNSIGNED4 day;
    UNSIGNED4 time;
    STRING11 Case_Number;
    STRING4 IUCR;
    STRING33 Primary_Type;
    STRING60 Description;
    UNSIGNED4 block_ID;
  END;

  EXPORT File := DATASET('~CLASS::HMW::OUT::Crimes_Slim', Layout, THOR);
END;
```

Solução proposta (BWR_Crime_Slim.ecl):

```
IMPORT $;

$.NormCrimeRecs.Layout Slimdown(RECORDOF($.Formatted_File) L,
                                $.NormAddrRecs.Layout R) := TRANSFORM
    SELF.block_ID := R.block_ID;
    SELF := L;

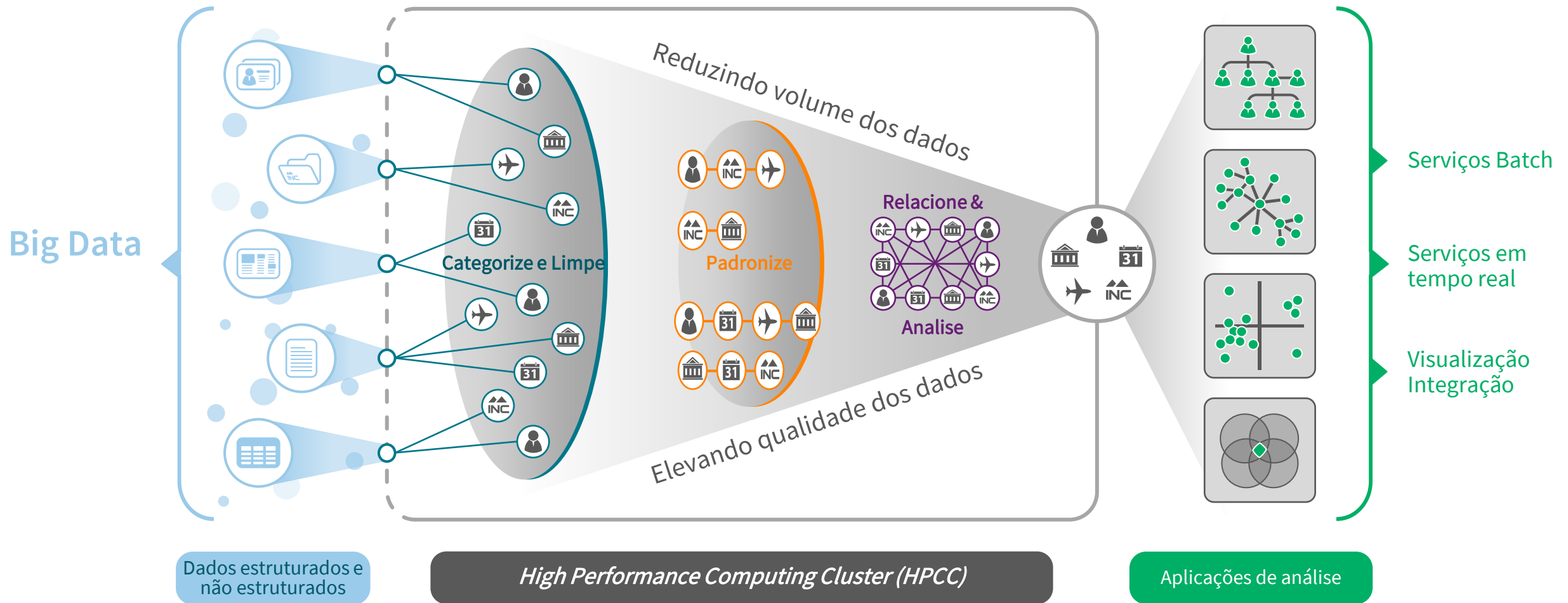
END;

SlimRecs := JOIN($.Formatted_File,$.NormAddrRecs.File,
    LEFT.block=RIGHT.block,
    Slimdown(LEFT,RIGHT),
    LEFT OUTER, LOOKUP);

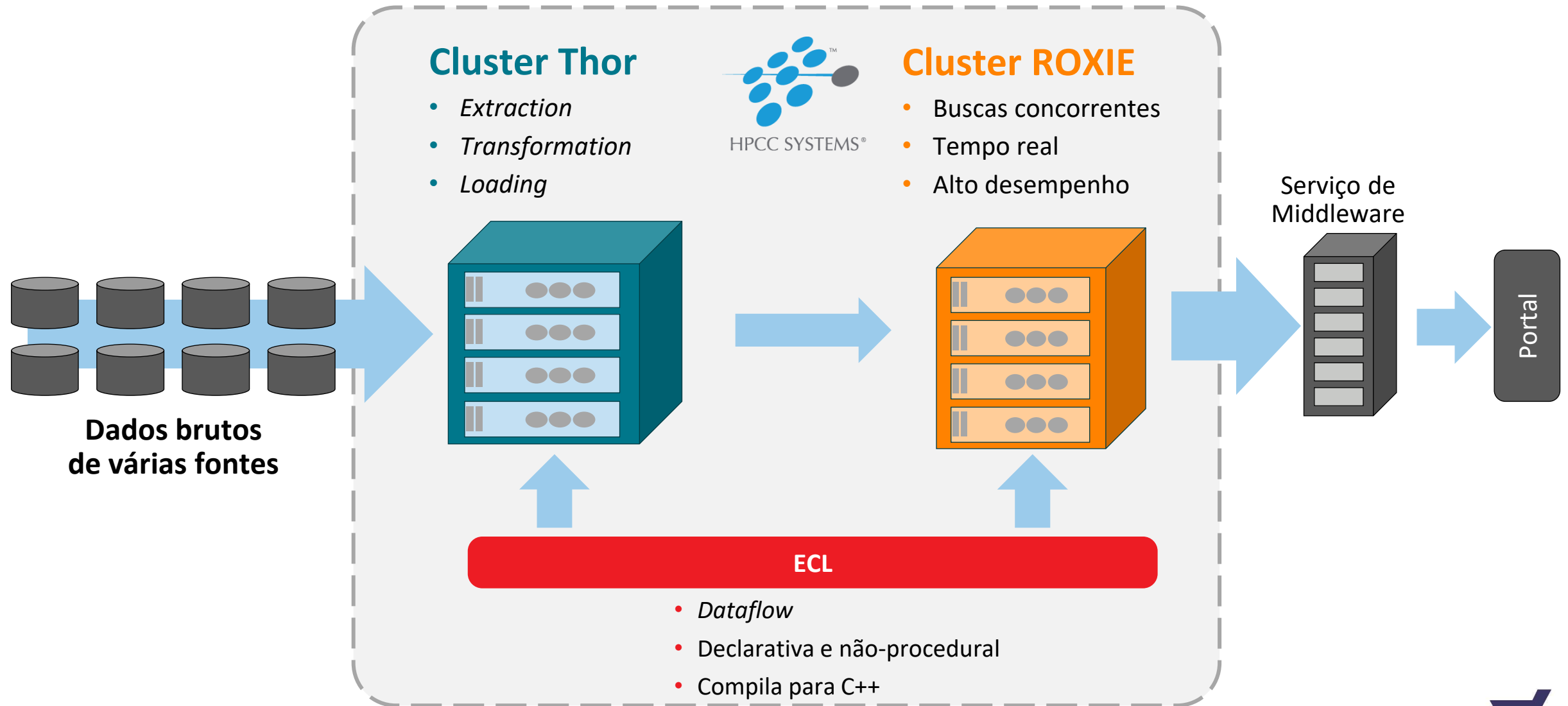
OUTPUT(SlimRecs,, '~CLASS::HMW::OUT::Crimes_Slim',overwrite);
```


Carregamento de dados

Extract, Transform, Load

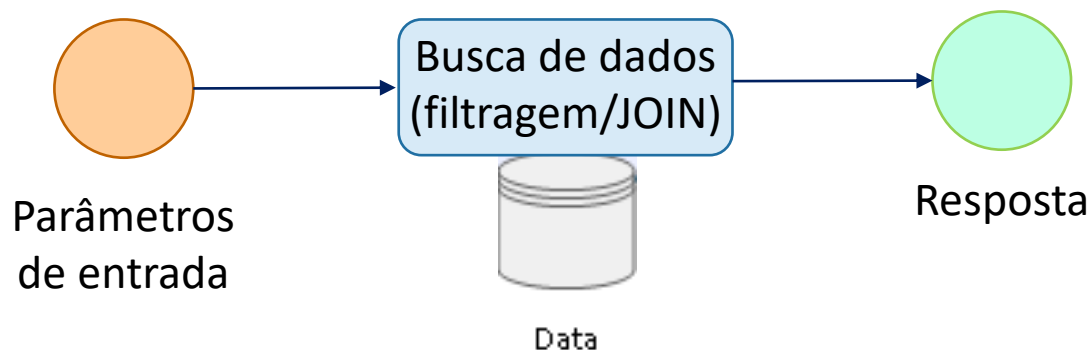


O power trio da plataforma: Thor, ROXIE e ECL



O que é uma query ROXIE?

- Método de busca otimizada a um conjunto específico de dados via uma solicitação web.



hthor

fn_fetchpersons-hmw

FN_FETCHPERSONS_HMWREQUEST

fname:

lname:

SMITH

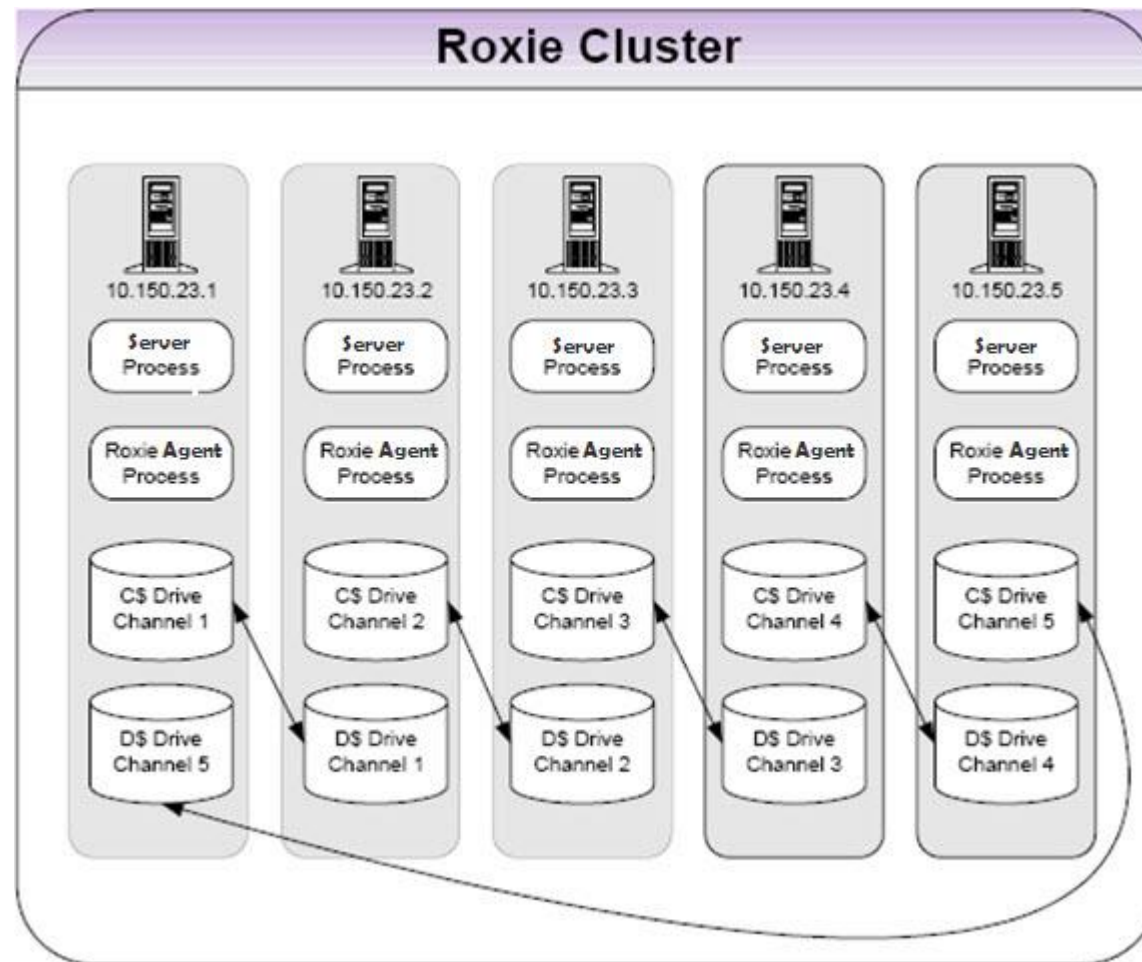
Output Tables

Dataset: Result 1

	lastname	firstname	recid	middlename	namesuffix	filedate	bureaucode	gender	dependentcount	birthdate	streetaddress	city	state	zipcode
1	SMITH	DETELIN	596872	M		19861027	171	U	0	19650301	338 W 89TH ST APT 3R	ARLINGTON	MN	55307
2	SMITH	CELENIA	644126			19920319	199	F	0	19221201	2355 FOREST HILLS DR	TRANSFER	PA	16154
3	SMITH	TEH	623354	L	SR	19871001	238	M	0	19561201	18 CROSS RIDGE RD	EAST ORANGE	NJ	7017
4	SMITH	YAMROT	341777			20000810	168	F	0	19611214	280 W 25TH ST # C	FAYETTEVILLE	NC	28314
5	SMITH	GANIJA	44886	Z		19870909	13	F	0	19260301	2090 POTTS HILL RD	POMPTON PLAINS	NJ	7444
6	SMITH	CASIANO	643584	N		19871210	78	F	0	19620219	1754 PALISADE AVE	PERU	IL	61354
7	SMITH	RUEI	727071	S		19950921	252	M	0	19751201	43 HILLAND DR	GLOUCESTER POIN	VA	23062
8	SMITH	ANNONAN	533969			19850821	376	F	0	19530108	134 E 17TH ST APT 63	JAMESON	MO	64647
9	SMITH	NAMIT	347861			19860401	24	M	0	19640928	221 CLERMONT AVE # 1	GLEN ELLYN	IL	60138
10	SMITH	MONTAKARN	277642	Q		19870309	13	M	0	19640929	45 MALLARD RD	DUNCANSVILLE	PA	16635
11	SMITH	VALDINA	609590	T		19940913	352	F	0	19730712	122 N BROAD ST	PITTSFIELD	NH	3263

O que é ROXIE?

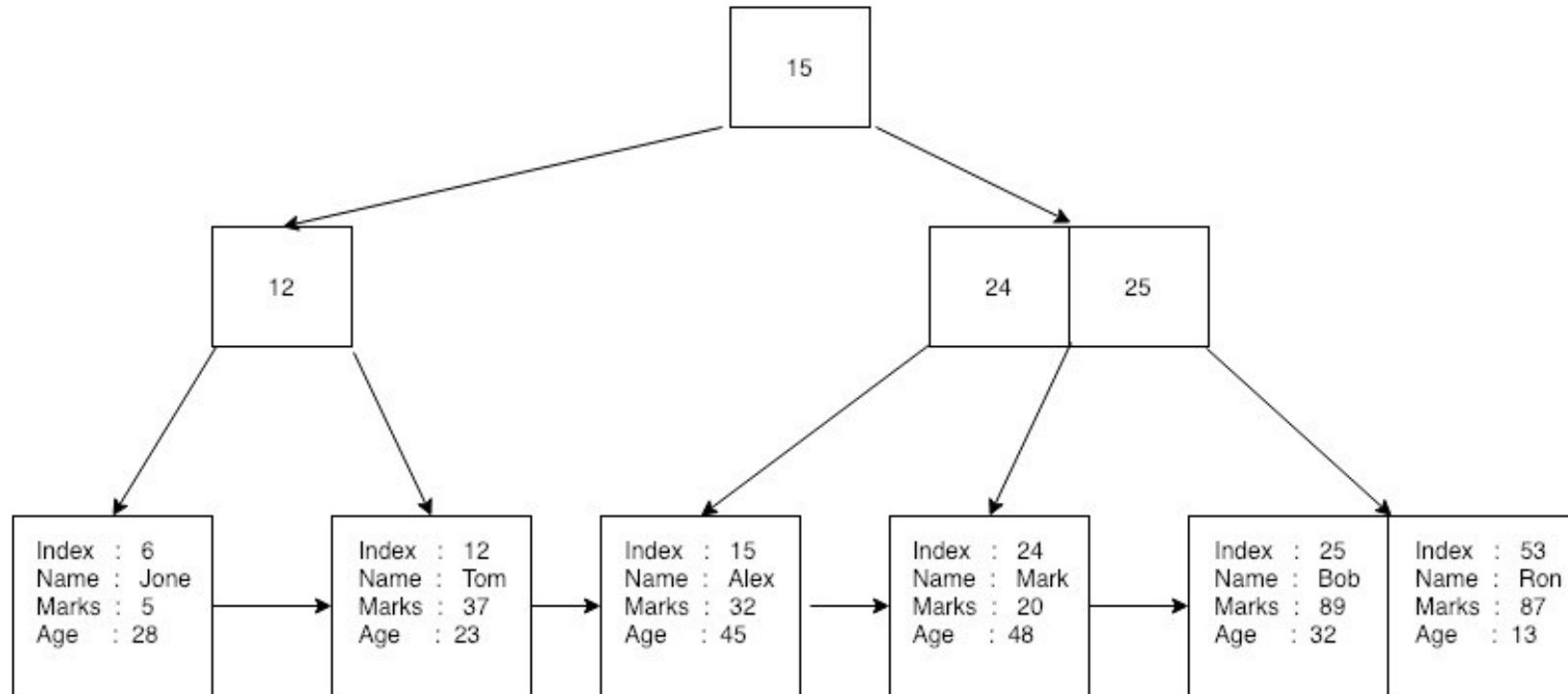
- ROXIE é um sistema de query massivamente paralelo, integrado ao HPCC.
- Conjunto de nós que:
 - Funcionam como uma única entidade que executam processos Servidores e os Agentes;
 - Executam múltiplas threads em cada nó para que os dados sejam recuperados de forma eficiente;
 - Utiliza índices com queries pré-compiladas;
 - Tempo de resposta em ms;



Exemplo de um cluster ROXIE de 5 nós.

Indexação de dados

- Índices possibilitam acesso rápido a um registro específico necessário para responder a uma consulta
- A construção de um índice envolve a ordenação dos dados



Índice no HPPC Systems

Arquivos Lógicos class::hmcw::key::lname_fname x

Recarregar

Abrir

Remover

Cópia Remota

Copiar

Renomear

A

☐

Nome Lógico

☐

class::hmcw::key::lname_fname

##	lastname i	firstname i	recid	id	middlename	namesuffix	filedate	bureaucode
1		ADAIR	368657	85871291...	G		19840201	127
2		DELOIS	193271	17719270...	D		19970130	238
3		NERMIN	786296	43851718...	Y		19860529	171
4		PAPA	500428	56792939...			19851012	209
5		RENURA	621153	50770447...	C		19961030	171
6		UNDINE	466812	17923479...			19900905	209
7	AAKJAR	BEERY	221637	14727661...	H		19930909	70
8	AAKJAR	DCASEY	455502	60055985...			19861117	6
9	AAKJAR	JENSON	468287	98406923...			19850330	406

Parte Cópia IP			Cluster (Aglome	Tamanho
1	1	10.0.0.210	mythor	15,785,984
1	2	10.0.0.175	mythor	15,785,984
2	1	10.0.0.175	mythor	15,302,656
2	2	10.0.0.228	mythor	15,302,656
3	1	10.0.0.228	mythor	15,663,104
3	2	10.0.0.210	mythor	15,663,104
4	1	10.0.0.210	mythor	40,960
4	2	10.0.0.175	mythor	40,960

Definição de Carregamento:

Disponibilização de dados para análises/consultas.

- Criação de índices no cluster THOR
- Publicação de dados e consultas para um cluster ROXIE

Como definir um índice?

A declaração **INDEX** define um arquivo de índice.

INDEX(*[base,] def, [payload,] arquivo [opções]*)

- ✓ *base* – O conjunto de registros a partir do qual o índice foi criado.
- ✓ *def* – A estrutura RECORD do arquivo de índice (campos de busca).
- ✓ ***payload*** – Uma estrutura **RECORD** com os campos **payload** (não serão campos de busca).
- ✓ *arquivo* – Nome do arquivo lógico contendo o arquivo de índice criado por um BUILD.
- ✓ *opções* – SORTED, PRELOAD, COMPRESSED, DISTRIBUTED.

```
EXPORT Key_Vehicle_City := INDEX( File_Vehicles,  
                                {st,city},  
                                {lname,fname},  
                                'key::PAY::vehicle.st.city');
```

Ação para construção do índice

A ação **BUILD** cria um índice por meio de uma definição INDEX.

BUILD(*indexdef*)

✓ *indexdef* – O nome da definição INDEX a ser criada.

nameKey := **INDEX**(Vehicles,{st,city},{lname},'vkey::st.city');

BUILD(nameKey);

Consulta: Função de busca

```
// FN_FindPerson.ec1
EXPORT FN_FetchPerson (STRING lname, STRING fname) := FUNCTION

    Person := $.File_Persons_Slim(LastName=lname AND FirstName=fname);

    RETURN OUTPUT(Person);
END;
```

```
//BWR_teste.ec1
OUTPUT(FN_FindPerson ('AAL','YARA'));
```

lastname <i>i</i>	firstname <i>i</i>	recid	id	middlename	namesuffi	filedate	bureaucode	gender	birthdate	streetaddress	csz_id
AAL	VERNDELL	465931	40928470...			19991119	2	F	0	225 E 74TH ST APT 2G	63388
AAL	XIANGEN	482886	83553240...	T		19931207	315	M	19590321	220 E 22ND ST	9624
AAL	YARA	819110	11703618...	N		19821229	171	M	19490722	368 CLARK ST	67975
AAL	YEWHALASHET	515962	93299555...			20000216	143	F	19791129	39 CEDAR AVE	17851
AALBERG	ANETT	190204	12305785...	N		19801001	240	F	19550901	340 E 64TH ST APT 32C	6861
AALBERG	BLADE	14962	15018122...	V		19860927	70	F	19450603	26 WILDWOOD ST	14962
AALBERG	CHARUNEE	393684	12014625...			19890503	393	F	0	5 VALLEY RD	31027

Estrutura FUNCTION

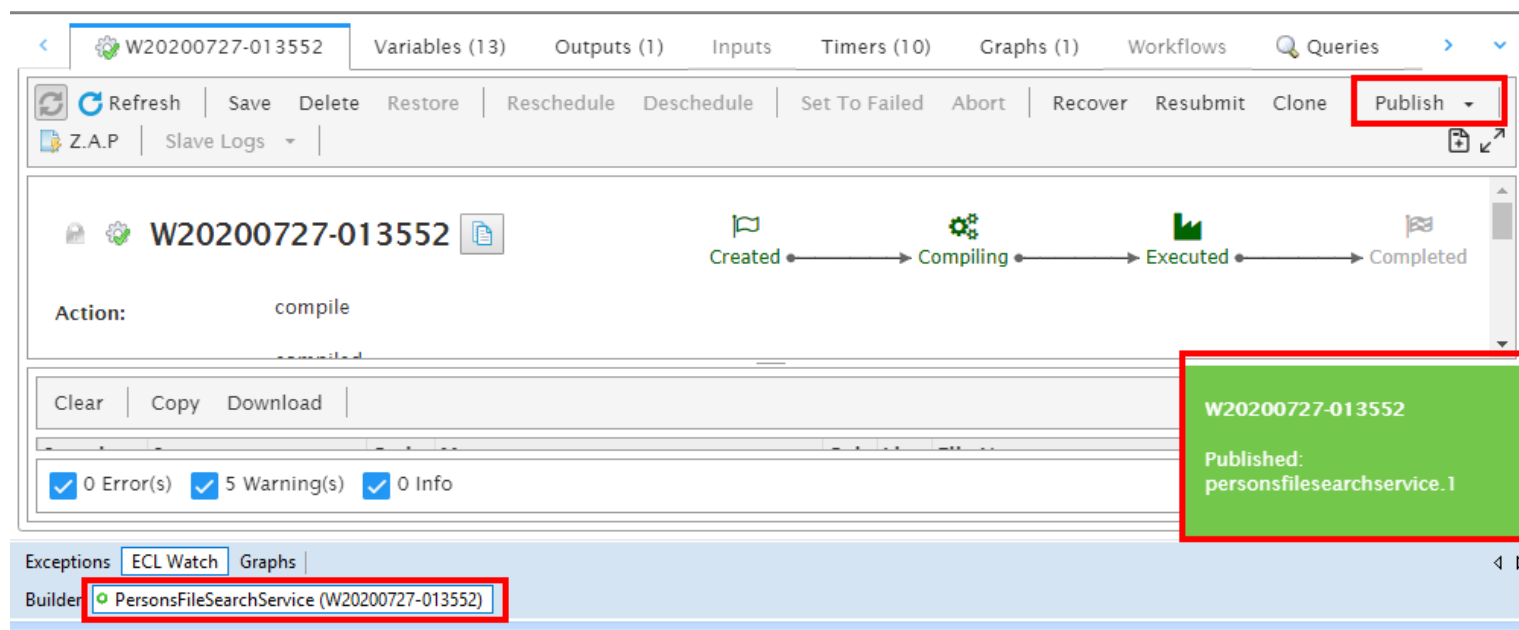
```
[tipo] nome (parametros) := FUNCTION  
    código;  
    RETURN retorno;  
END;
```

- *tipo* – O tipo do valor de retorno da função.
- *nome* – O nome da definição ECL para a função.
- *parametros* – Os parâmetros a serem fornecidos ao *código*.
- *código* – As definições no processos da função.
- *retorno* – O valor, expressão, recordset, linha (registro) ou ação a ser retornada.

A estrutura **FUNCTION** permite o fornecimento de parâmetros comuns para um conjunto de funções relacionadas.

Publicação da consulta

- Depois de elaborar a sua consulta, é necessário publicá-la no cluster ROXIE:
 - 1) Selecione como alvo o cluster ROXIE no seu ECL IDE
 - 2) Clique “Submit > Compile”
 - 3) Abra a WU compilada no ECL Watch e clique em “Publish”



Teste a consulta publicada

- ECL Watch > Queries

The screenshot shows the ECL Watch web interface. At the top, there's a blue header with the 'ECL Watch' logo and several icons. Below the header, there's a navigation bar with tabs for 'Queries' and 'Package Maps'. The 'Queries' tab is active, and a sub-tab for 'personsfilesearchservice.1' is selected. Below this, there's a row of tabs: 'Summary', 'Errors/Status (1)', 'Logical Files (5)', 'Super Files', 'Libraries Used (0)', 'Graphs (1)', 'Resources', 'Test Pages', and 'W20200727-013552'. The 'Test Pages' tab is highlighted with a red box. Below this, there's another row of tabs: 'SOAP', 'JSON', 'WSDL', 'Request Schema', 'Response Schema', 'Sample Request', 'Sample Response', 'Parameter XML', 'Legacy Form', and 'Links'. The 'Legacy Form' tab is also highlighted with a red box. The main content area shows a 'Reset' button and a section for 'roxie' with a dropdown menu set to 'Dynamic Form'. Below this, there's a section for 'PERSONSFILESEARCHSERVICE_1REQUEST' with a checked checkbox. This section contains four input fields: 'firstname:', 'lastname:', 'sex:', and 'state:'. A red arrow points to the 'PERSONSFILESEARCHSERVICE_1REQUEST' section. At the bottom, there's a row of controls: a 'Call Query' dropdown, an 'Output Tables' dropdown, a 'FORM POST' dropdown, a 'Submit' button (highlighted with a red box), and a 'Clear All' button. There's also a 'Capture Log Info.' checkbox and a 'Trace Level:' input field.

Teste a consulta publicada

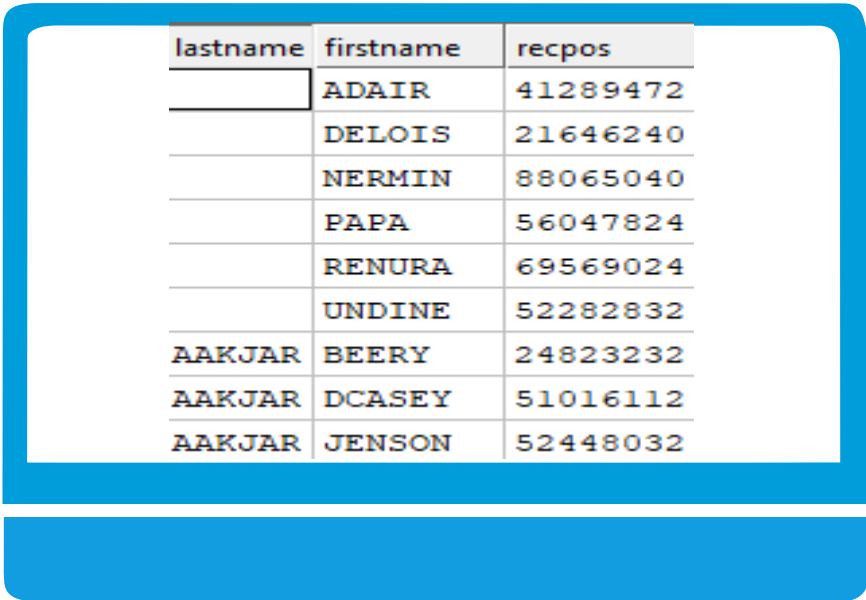
- WsECL (porta 8002):

The screenshot shows the HPCC Systems WsECL 3.0 web interface. The browser address bar shows 'play.hpccsystems.com:8002'. The interface has a top navigation bar with 'HPCC Systems', 'View', 'Frame', 'Log Out', and 'WsECL 3.0'. Below this is a sidebar with 'Active Queries' and a tree view of targets: 'thor', 'hthor', 'roxie' (highlighted with a red box), and 'thor_roxie'. The 'roxie' target is expanded, showing 'personsfilesearchservice'. The main area displays the 'personsfilesearchservice' form, which is a 'Dynamic Form'. The form title is 'PERSONSFILESEARCHSERVICEREQUEST'. It contains four input fields: 'firstname:', 'lastname:', 'sex:', and 'state:'. A red arrow points to the 'lastname:' field. At the bottom, there is a 'Capture Log Info.' checkbox, a 'Trace Level:' dropdown, and a row of buttons: 'Call Query', 'Output Tables', 'FORM POST', 'Submit' (highlighted with a red box), and 'Clear All'.

Exercício prático

Exercícios 1a-1d – Criando índices (Persons_Slim e Lookup_CSZ)

- Utilize MODULE original
- Exponha campo filepos
- Definição INDEX
- Ação BUILD



lastname	firstname	recpos
	ADAIR	41289472
	DELOIS	21646240
	NERMIN	88065040
	PAPA	56047824
	RENURA	69569024
	UNDINE	52282832
AAKJAR	BEERY	24823232
AAKJAR	DCASEY	51016112
AAKJAR	JENSON	52448032

Exercício prático

Exercícios 1e-1f – Função e serviço de busca

- Estrutura FUNCTION (Persons_slim e Lookup_CSZ)
- JOIN como retorno
- Serviço STORED
- Publicação da consulta

personsfilesearchservice.1 Response

Dataset: Result 1





	recid	id	firstname	lastname	middlename
1	596872	1166595353551052552	DETELIN	SMITH	M
2	644126	5082066557802196041	CELENIA	SMITH	
3	623354	17588373756131785978	TEH	SMITH	L
4	341777	8451032049636405291	YAMROT	SMITH	
5	44886	7830553978810542989	GANIJA	SMITH	Z
6	643584	1292861395782975695	CASIANO	SMITH	N
7	727071	12221818927523640828	RUEI	SMITH	S
8	533969	17844693484560518926	ANNONAN	SMITH	
9	347861	675797059124681114	NAMIT	SMITH	
10	277642	4364131847069821064	MONTAKARN	SMITH	Q
11	609590	5832249607040579389	VALDINA	SMITH	T

Desafio: Chicago Crimes

Desafio Chicago Crimes:

- Crime um serviço de busca de crimes

roxie

desafio7_crimesvc-short.1     Dynamic Form ▾

DESAFIO7_CRIMESVC_SHORT_1REQUEST ☒

block_info:

☐ Capture Log Info. Trace Level: ☐ No Timeout

Call Query ▾ Output Tables ▾ FORM POST ▾ Submit Clear All

Dataset: Result 1

	block	primary type	cnt
1	0000X S STATE ST	THEFT	817
2	001XX S STATE ST	THEFT	284
3	0000X S STATE ST	DECEPTIVE PRACTICE	274
4	008XX S STATE ST	THEFT	230
5	002XX S STATE ST	THEFT	222
6	098XX S STATE ST	CRIMINAL DAMAGE	191
7	005XX S STATE ST	THEFT	187
8	011XX S STATE ST	THEFT	154
9	004XX S STATE ST	THEFT	138
10	002XX S STATE ST	BATTERY	81

Quais foram os objetivos principais desse curso?

- ✓ Familiaridade com o processo de ETL
 - ✓ Extração (Spray/Profiling)
 - ✓ Transformação (Normalização)
 - ✓ Carregamento (Publicação de consultas)
- ✓ Compreensão de conceitos e sintaxe em ECL
- ✓ Domínio de quatro fundamentos básicos de ETL em ECL:
 - ✓ Compreenda seus dados
 - ✓ Opere somente nos dados necessários
 - ✓ Transforme os dados para menor ocupação de armazenagem possível
 - ✓ Faça uso de estratégias que otimizem seu processo

Próximos passos

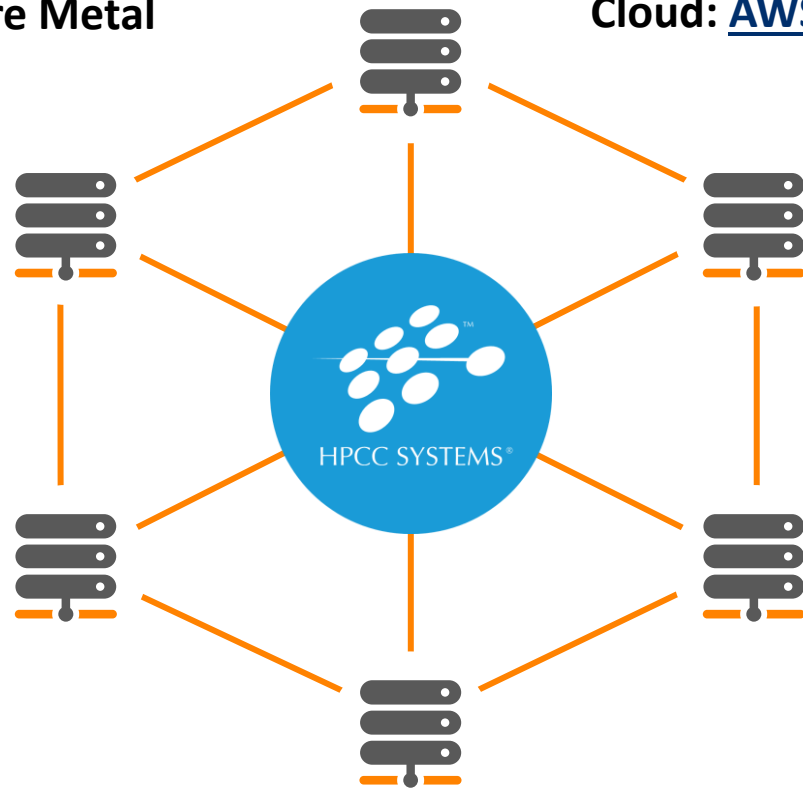
Próximos passos

- ✓ Enviar badges de certificação e exemplos de Código
- ✓ Certificado USP
- ✓ Conferência HPCC Systems (Outubro/22)
 - ✓ <https://hpccsystems.com/community/events/hpcc-systems-summit-2022>
- ✓ Playground / Treinamento online / documentação / fórum

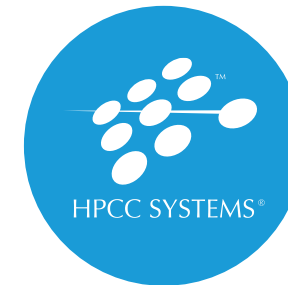
Opções de uso: play.hpccsystems.com

Bare Metal

Cloud: [AWS](#)/[Azure](#)



Oracle Virtual Box
HyperV
[Docker](#)
GitPod



[HPCC Máquina Virtual](#)

✓ <https://hpccsystems.com/pt-br/try-now>

Cursos online: +170 aulas (learn.lexisnexus.com/hpcc)

- Introdução ao ECL (parte 1)
 - Conceitos e consultas
- Introdução ao ECL (parte 2)
 - ETL com ECL
- ECL Avançado (parte 1)
 - Dados relacionais
- ECL Avançado (parte 2)
 - Superarquivos, XML/JSON e PLN
- ECL Aplicado
 - Geração e automação de código ECL

ROXIE ECL (parte 1)

- Índices e consultas

ROXIE ECL (parte 2)

- Otimização de consultas

Machine Learning com HPCC Systems

- Fundamentos para uso dos plugins

Administração de Sistemas

- Conceitos e operação básica

HPCC para gestores

- Visão geral e aplicações da plataforma

Benchmark

Table 1. HPCC vs Hadoop vs Spark

Topic	HPCC	Hadoop	Spark
Parallelism Paradigm	Dataflow Three parallel execution modes: <ul style="list-style-type: none">• Data: Data partitioned across nodes; Compute occurs on each node in parallel• Pipeline: Consecutive operations on the same dataset at the same time; Data processed by one operation immediately passed to the next• System: Independent operations try to execute in parallel	MapReduce Data parallelism only, and only in the Map phase.	RDD (Resilient Distributed Dataset) Data parallelism only

Topic	HPCC	Hadoop	Spark
Compilation	Yes. The C++ generated by the ECL Compiler is compiled for execution	No. JVM-based	No. JVM-based
Built-in End User Query Support	Yes. Roxie clusters deliver thousands of concurrent end-user transactions per second (actual numbers dependent on the number of nodes in the cluster and the complexity of the queries themselves)	No. Third party tools required.	No. Third party tools required.
Production Monitoring	Yes. Ganglia and Nagios included as part of the platform.	No. Third party tools required.	No. Third party tools required.
Language(s) Supported	ECL built in with any other language embeddable inline. C++, Java, Javascript, Python, SQL, and R currently supported. More embedded languages can be added by the community	Java, Hive, Pig	API allows JVM-based language programming (like Java, Python, Scala, and R)

https://cdn.hpccsystems.com/whitepapers/hpccsystems_thor_spark.pdf

Relacionamento com Academia

<https://hpccsystems.com/community/academics>



Universidade de São Paulo
Brasil



UNIVERSIDADE FEDERAL
DE SANTA CATARINA



Projetos de Pesquisa



<https://wiki.hpccsystems.com/display/hpcc/Available+Projects>

Links úteis

- Site principal: hpccsystems.com
- Primeiros passos: hpccsystems.com/Why-HPCC-Systems
- Canal do youtube: youtube.com/user/HPCCSystems
- Fórum da Comunidade: hpccsystems.com/forums



Faça parte da Comunidade

Registre-se em hpccsystems.com

O Grupo RELX



Reed Elsevier 1992 – 2015
Reed Internacional (1894) + Elsevier NV (1880) - Editoras



Risk & Business Analytics - Serviços de dados e tecnologia, análises, informações preditivas e prevenção de fraudes para uma ampla gama de setores.

		BR
	Análise de risco (BS, Saúde, Jurídico, Governo, Seguros)	130
	Verificação de ativos financeiros, meios de pagamento, crimes financeiros	11
	Informação de preços de commodities (petroquímicos)	7
	Análise de dados para setor aeroespacial	
	Gestão de informação, análise de dados e compliance de Recursos Humanos	
	Produtos e serviços de dados, notícias e análises para o mercado imobiliário comercial.	
	Conectividade e gestão de dados para as indústrias de agricultura e saúde animal.	5
	Soluções fiscais, conectividade e declaração online.	

Até o próximo curso!!!

