Dominando Big Data con HPCC Systems

Descripción general y aplicaciones de la plataforma

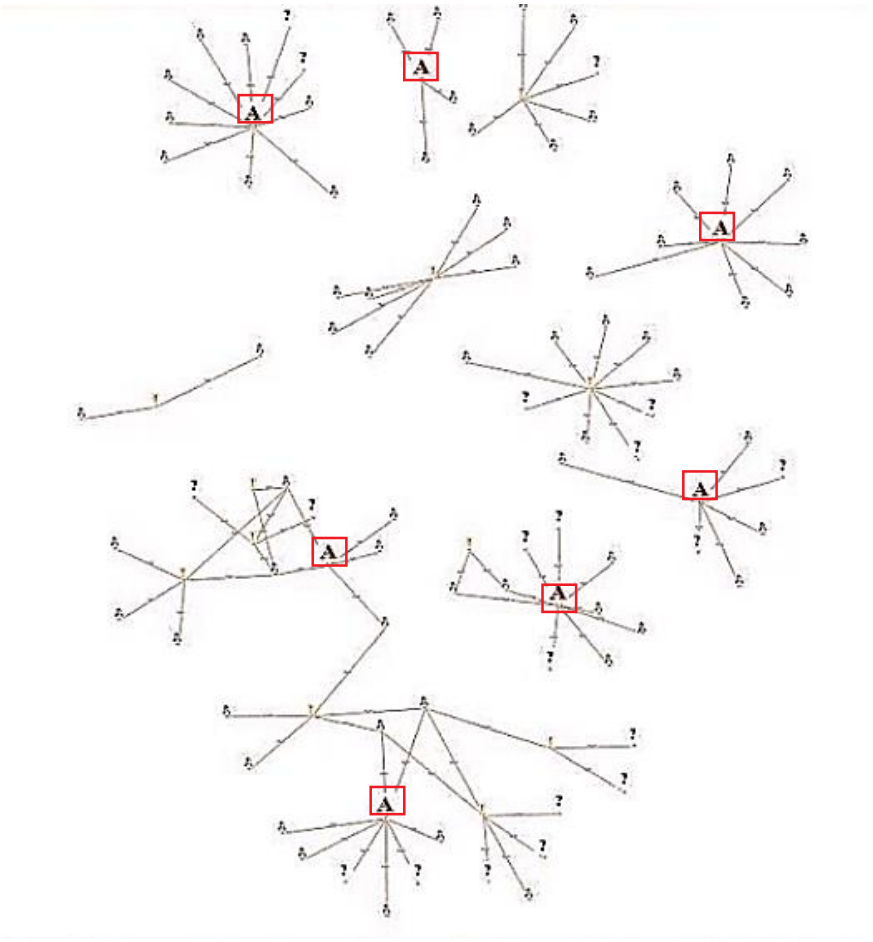
HPCC SYSTEMS®

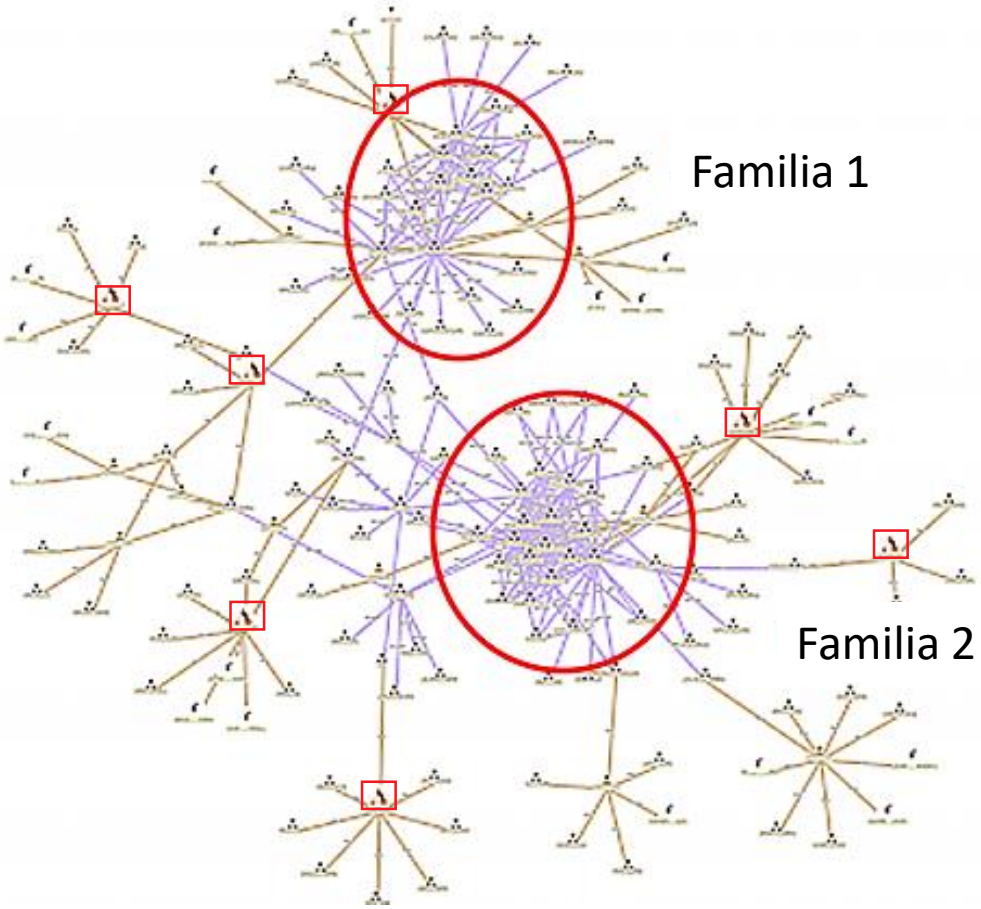
Universidad Católica de Pereira
Facultad de Ciencias Económicas y Administrativas

LexisNexis® RISK SOLUTIONS

RELX Group

Caso de uso: Fraude de seguros





Accidentes con sospecha de fraude

Personas asociadas a accidentes

HPCC SYSTEMS®

¡Bienvenido! - Agenda del taller

- ✓ HPCC Systems: descripción general
  - ✓ ¿Qué es? ¿De donde vino? ¿Para que sirve?
- ✓ Tutorial: ETL con ECL
  - ✓ Hands-on
- ✓ Bundles y aplicaciones
  - ✓ Visualización
  - ✓ PLN
  - ✓ Machine Learning

HPCC SYSTEMS®

Recursos del taller

- Computadora personal (ECL IDE v.7.6.64 / VSCode)
- Clúster HPCC Systems: http://3.139.124.33:8010/
- Diapositivas y códigos (github.com/hpccsystems-solutions-lab/hpcc-systems-BR)
- Certificado de participación

HPCC SYSTEMS®

¿Quienes somos?

- ✓ hugo.watanuki@lexisnexisrisk.com
- ✓ robert.foreman@lexisnexisrisk.com
- ✓ richard.taylor@lexisnexisrisk.com
- ✓ https://hpccsystems.com/bb/

Meet the Trainers

Richard Taylor

Richard is the original author of the ECL documentation, developer and designer of the HPCC Systems Training Courses, and is the Chief Instructor for all classroom and remote based training.

Bob Foreman

Bob is the developer and designer of the HPCC Systems Online Training Courses, and is the Senior Instructor for all classroom and online based training.

Hugo Watanuki

Hugo supports the development and delivery of training programs for the HPCC Systems platform in the Brazil region.

HPCC SYSTEMS®

HPCC Systems: descripción general

HPCC SYSTEMS®

¿Qué es?

HPCC Systems (*High Performance Computing Cluster*) es una plataforma para resolver desafíos de Big Data:

- **Supercomputación**: procesamiento paralelo y datos distribuidos
- **Open source:** libre y de código abierto
- **Completa:** gestión de flujo de datos completa y simplificada

HPCC SYSTEMS®

¿De donde vino?

2001

Se lanza la primera versión

2011

Código abierto (licencia Apache y código en GitHub)

2012 – 16

Mejoras continuas con foco en la calidad

Capacitación y soporte mejorados

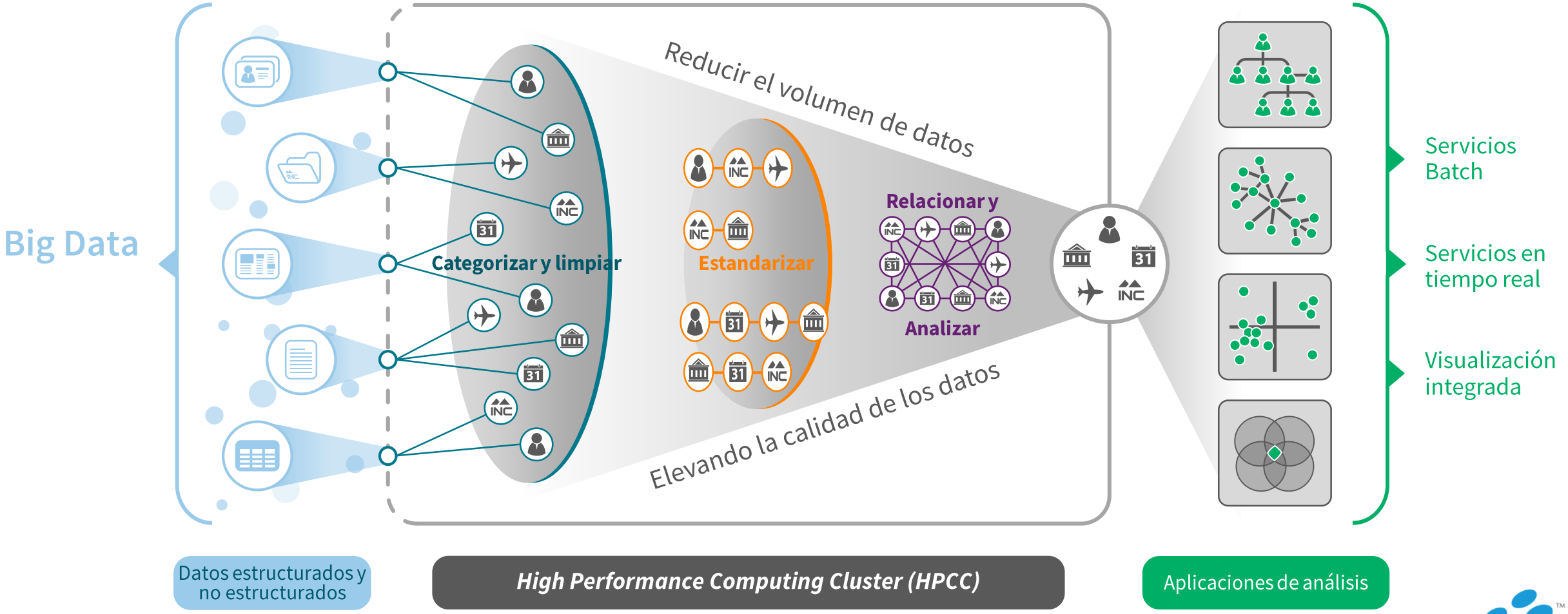
2017-actualidad

Mejoras arquitectónicas (Cloud)

Desarrollos en Machine Learning

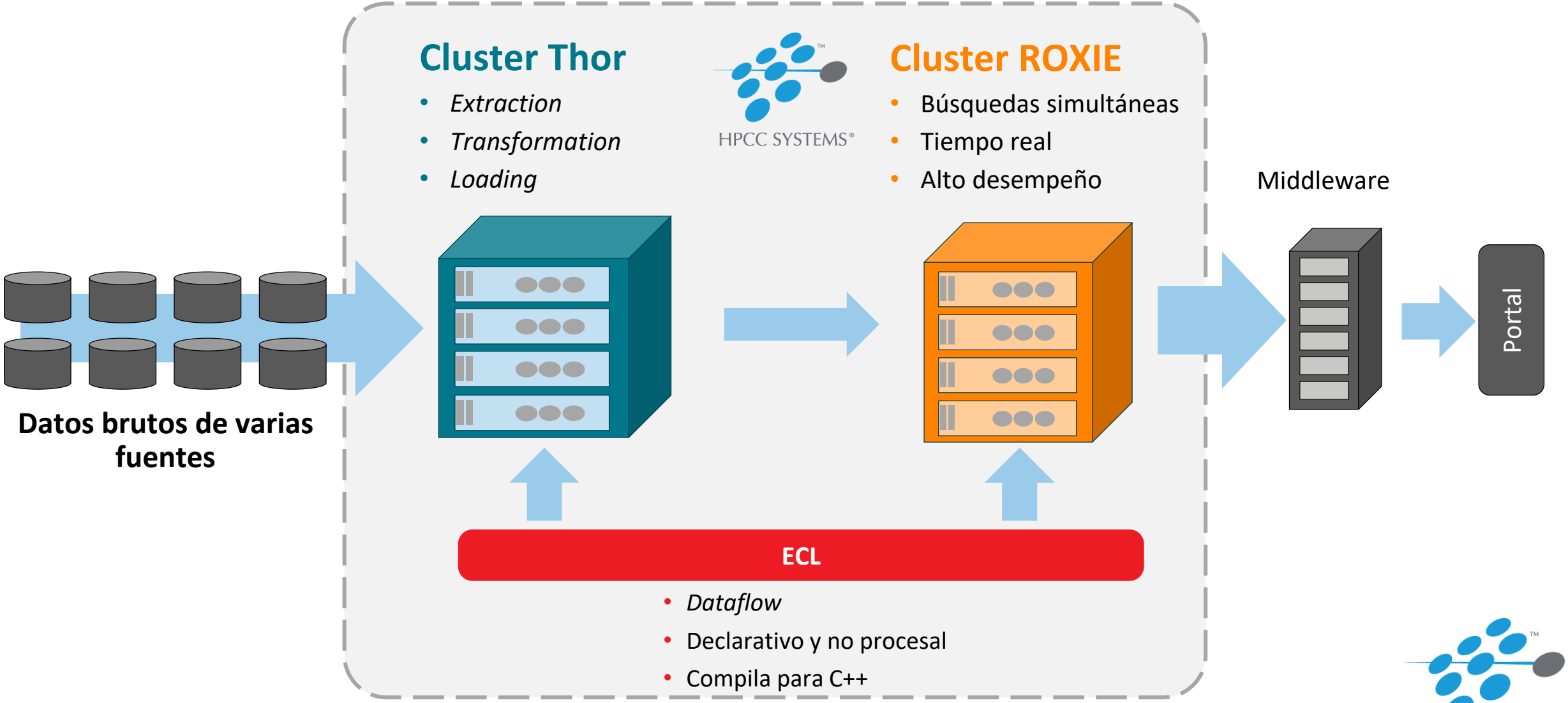
HPCC SYSTEMS®

¿Para que sirve?



HPCC SYSTEMS®

El *power trio* de la plataforma: Thor, ROXIE y ECL



HPCC SYSTEMS®

Thor y ROXIE: objetivos específicos

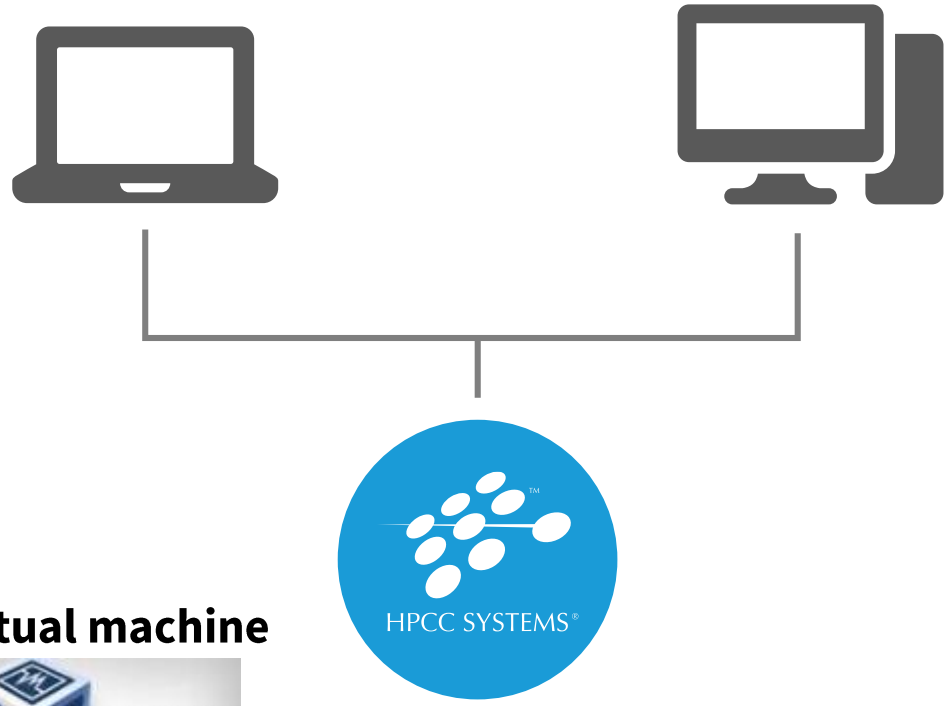
Thor:
"Identificar y catalogar todos los seres vivos de los océanos."

ROXIE:
"Poner a disposición toda la información sobre una especie"

HPCC SYSTEMS®

La plataforma puede funcionar en ...

… una sola computadora.

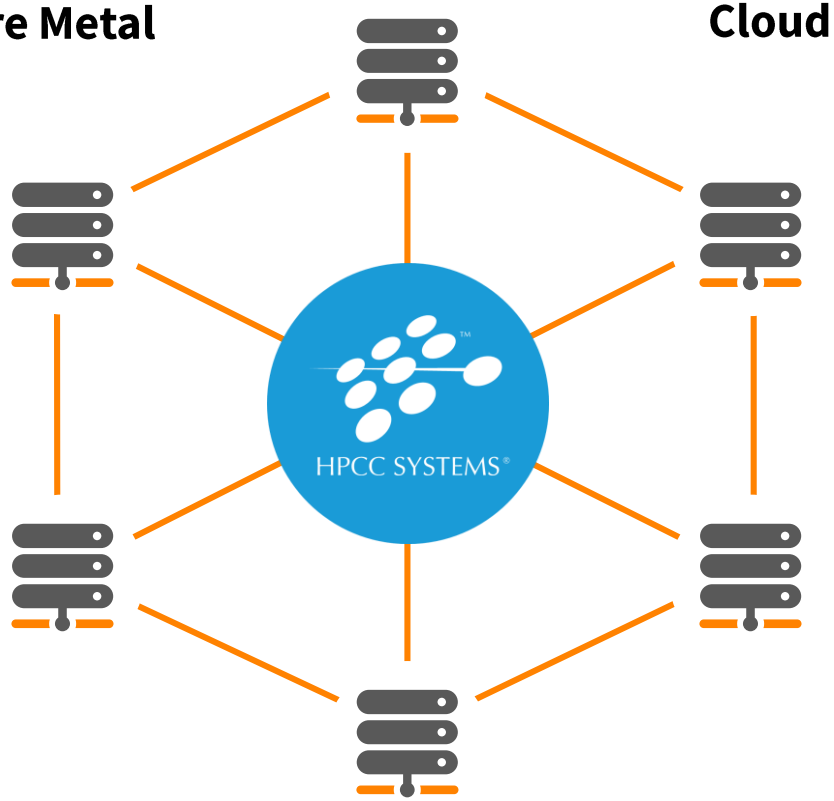


Virtual machine

… un clúster.

Bare Metal

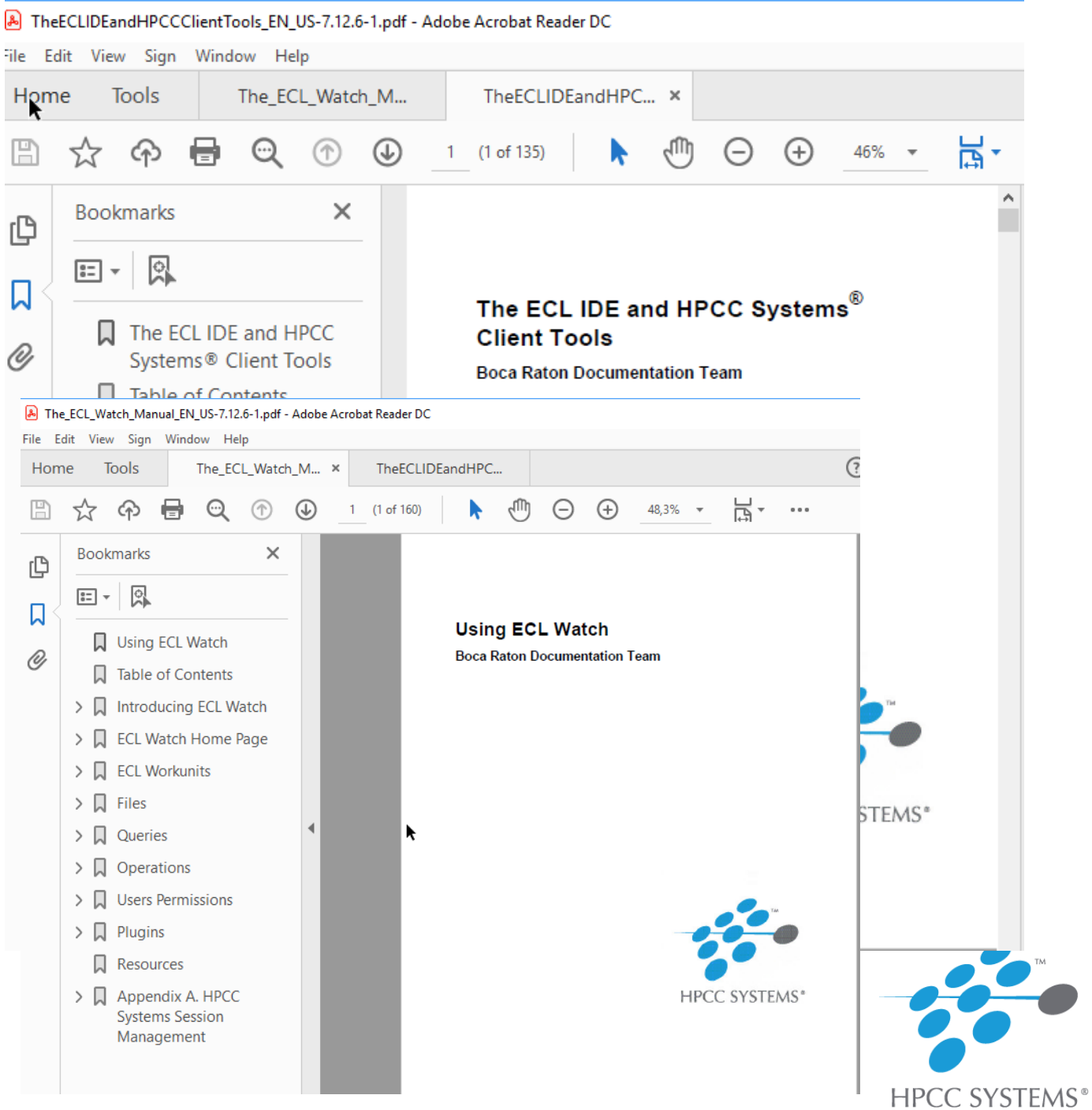
Cloud



HPCC SYSTEMS®

Interfaces de uso

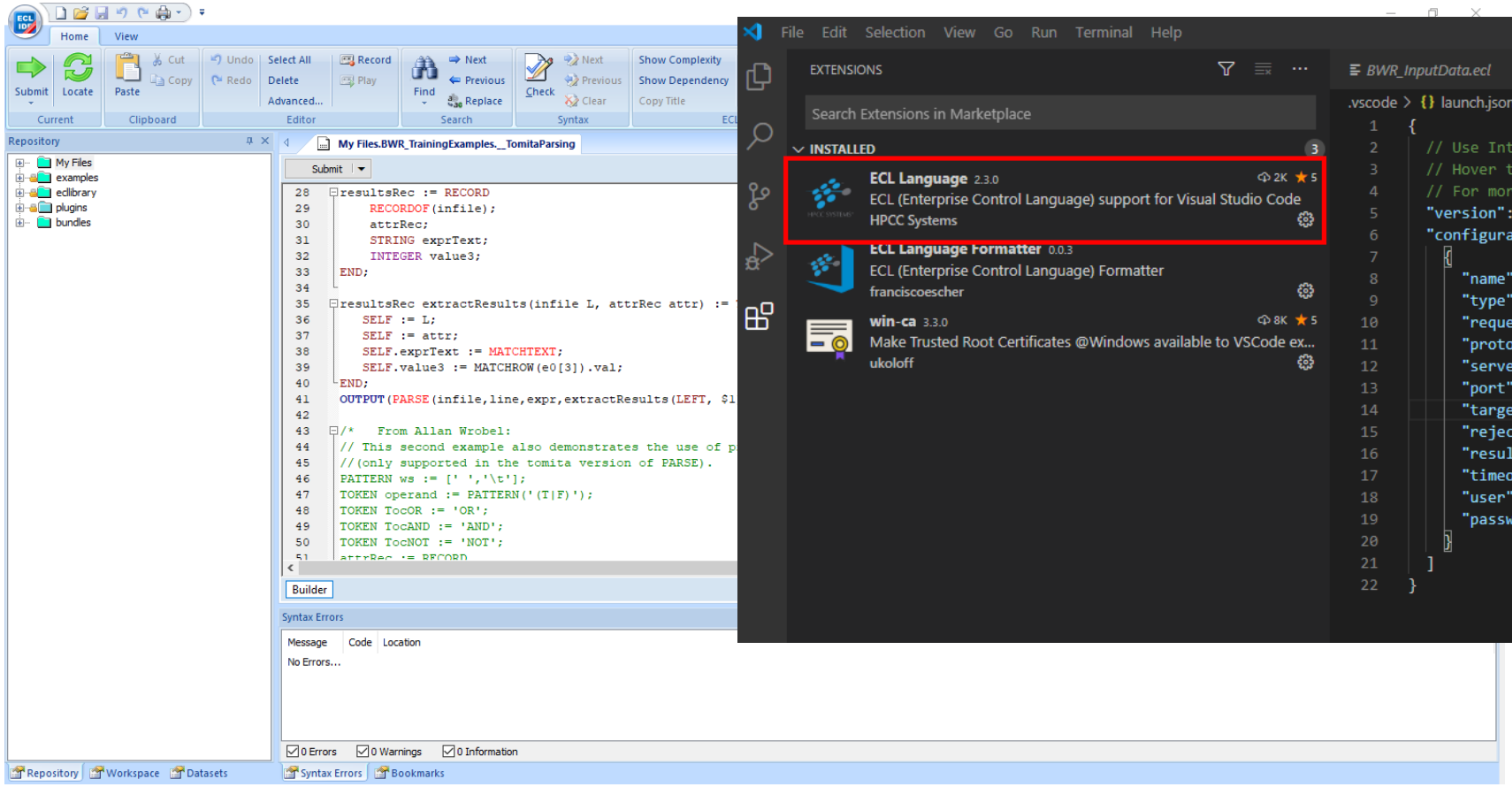
- ✓ ECL IDE
  - ✓ Herramienta de desarrollo ECL
- ✓ ECL CLI
  - ✓ Interface de línea de comandos
- ✓ ECL Watch
  - ✓ Herramienta web de gestión y supervisión

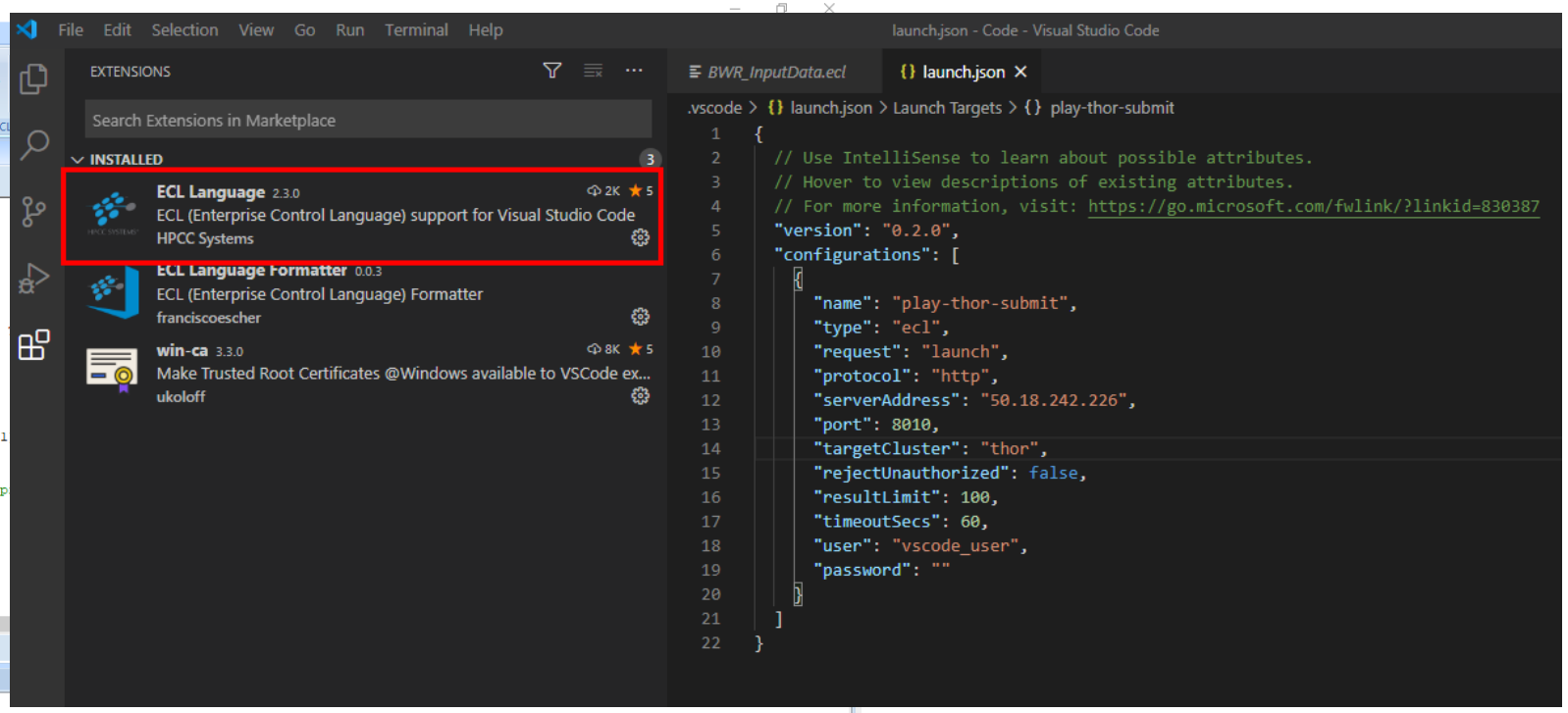


ECL IDE

✓IDE (Win)

✓VSCode (Ux/MacOS)





HPCC SYSTEMS®

ECL Watch

✓ Interface web (<ip>:8010)

ECL Watch

Wuid, User, (ecl:*, file:*, dfu:*,

LOGGED IN AS:

Activity | Event Scheduler | Search Results

Activity

Refresh | Pause | Resume | Clear | Open | Pause | Pause Now | Resume | Abort | High | Normal | Low | Top | Up | Down | Bottom

46% hthor

46% thor

28% roxie

46% thor_roxie

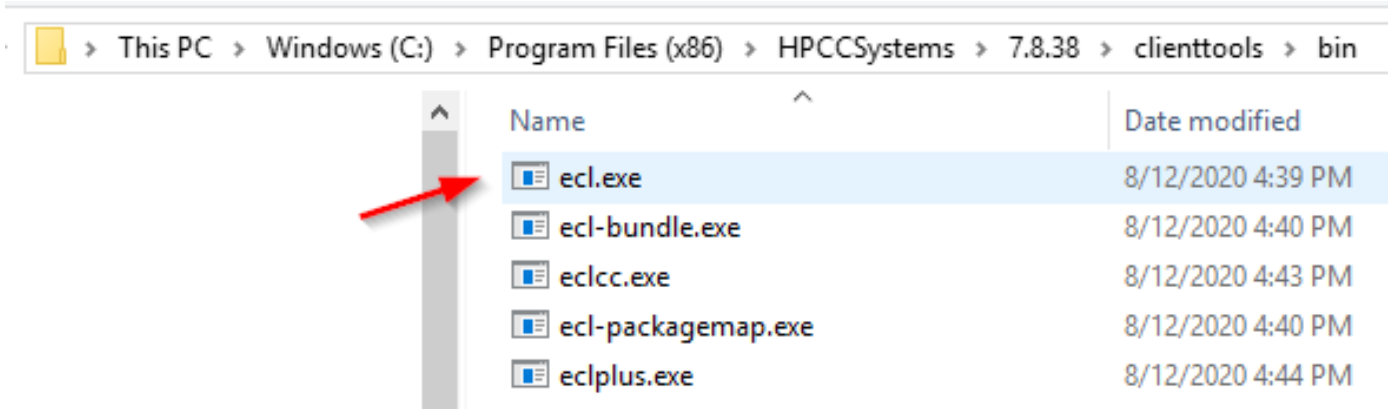
Target/Wuid	Graph	State	Owner	Job Name
HThorServer - hthor				
ThorMaster - thor				
ThorMaster - thor_roxie				
RoxieServer - roxie				
myeclccserver - hthor.eclserver				
myeclccserver - thor.eclserver				
myeclccserver - roxie.eclserver				
myeclccserver - thor_roxie.eclserver				
mydfuserver - dfuserver_queue				

HPCC SYSTEMS®

ECL CLI

✓ Clienttools (Win/Unix)

```
hugo@hugo-VirtualBox:/opt/HPCCSystems/bin$ ls -alrt ecl*
-rwxr-xr-x 1 root root   26776 set 23 14:43 eclscheduler
-rwxr-xr-x 1 root root 4016848 set 23 14:43 ecl-roxie
-rwxr-xr-x 1 root root 5970920 set 23 14:43 ecl-queries
-rwxr-xr-x 1 root root 5958640 set 23 14:43 eclplus
-rwxr-xr-x 1 root root 1169568 set 23 14:43 ecl-packagemap
-rwxr-xr-x 1 root root   67736 set 23 14:43 eclccserver
-rwxr-xr-x 1 root root  236720 set 23 14:43 eclcc
-rwxr-xr-x 1 root root 1449272 set 23 14:43 ecl-bundle
-rwxr-xr-x 1 root root 6142992 set 23 14:43 ecl
hugo@hugo-VirtualBox:/opt/HPCCSystems/bin$
```



HPCC SYSTEMS®

Fin de la parte 1!

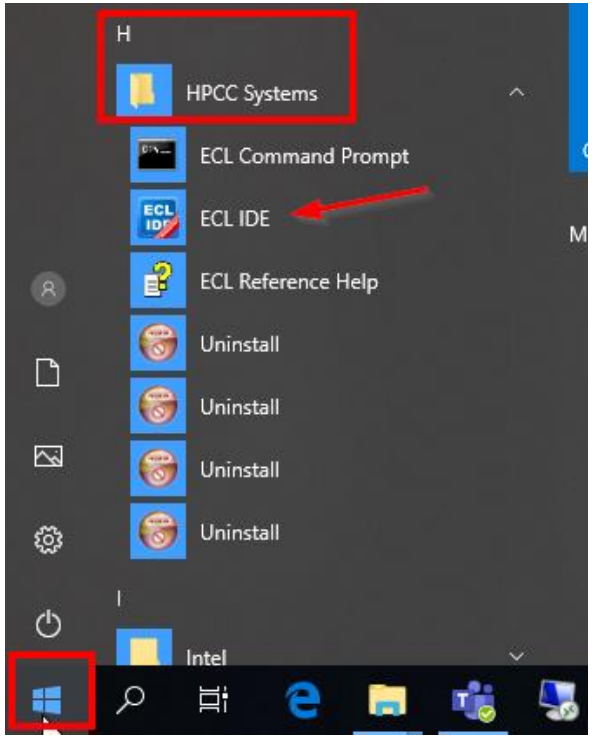
HPCC SYSTEMS®

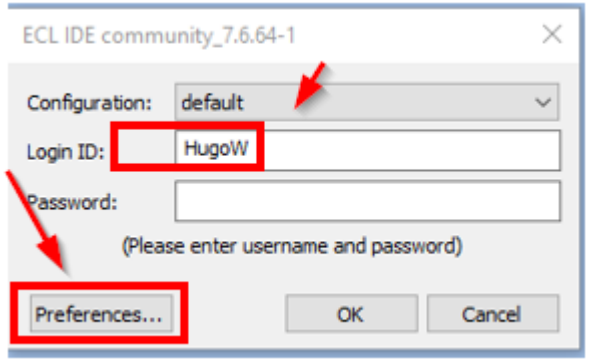
Tutorial: ETL con ECL

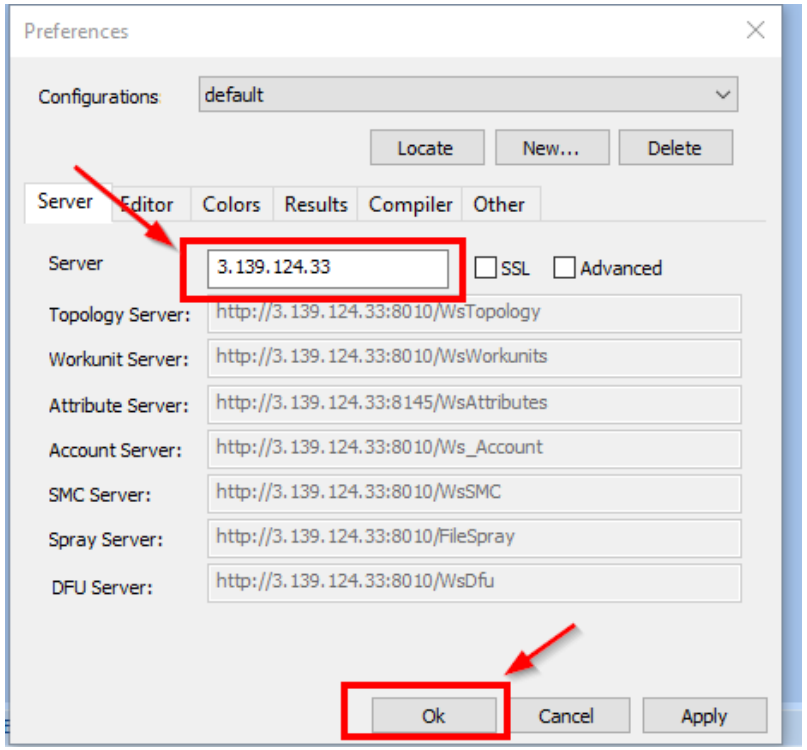
HPCC SYSTEMS®

Preparación del entorno

- Clúster: http://3.139.124.33:8010/
- ECL IDE:



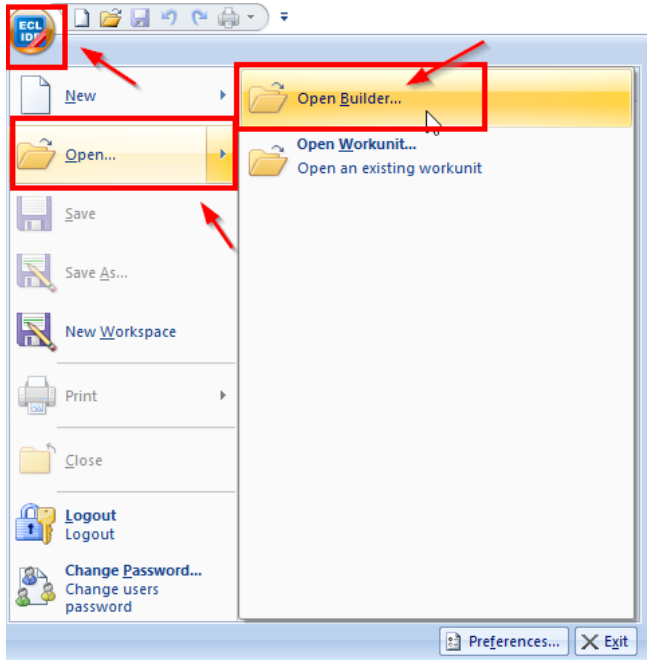


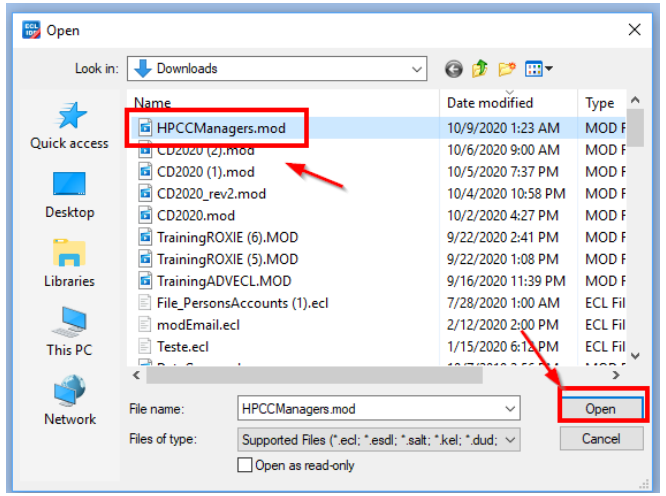


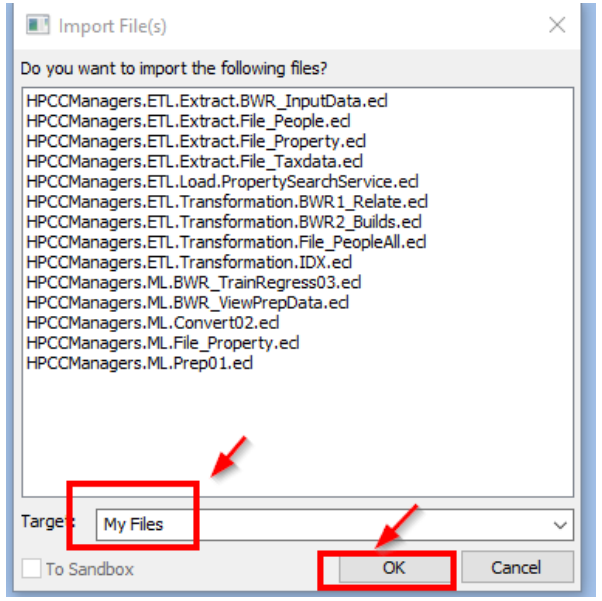
HPCC SYSTEMS®

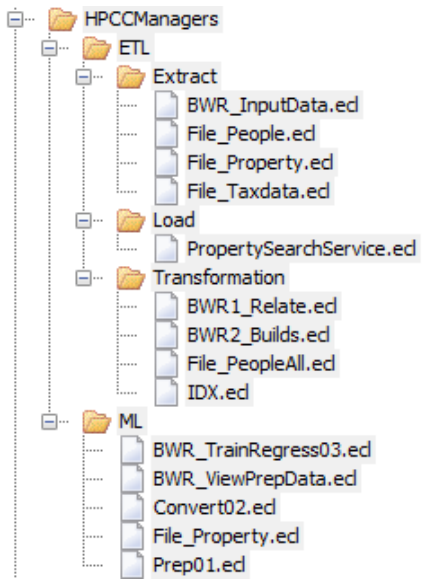
Preparación del entorno (cont.)

- HPCCManagers.mod









HPCC SYSTEMS®

ECL: Hello World!

*My Files.CodeCamp.BWR_MyFirstCodeProject

Submit

Target: thor

```
// EXPORT BWR_MyFirstCodeProject := 'todo';

// Define an attribute with a string literal
MY_STRING := 'Hello World!';

// Action that displays the string in the workunit
OUTPUT(MY_STRING);
```


HPCC SYSTEMS®

Enterprise Control Language (ECL)

- Lenguaje de programación centrado en datos (flujo de datos)
  - Declarativa y no procesal
  - Códigos reutilizables más pequeños
  - Biblioteca para manipulación de datos

- Compilador
  - Genera código optimizado (C++)
  - Lógica para procesamiento paralelo y distribuido

Como hacer ✗ vs. ✓ Que hacer

HPCC SYSTEMS®

Conceptos básicos de ECL

- Estructura basica : **`Nombre := Expressión;`**
- ECL no distingue entre mayúsculas y minúsculas
- Se ignoran los espacios en blanco para una mejor lectura
- Comentarios en línea (//) y en bloque (/* y */)
- ECL utiliza la sintaxis de objetos

  Dataset.Campo // hace referencia a un campo en un conjunto de datos

  NombreDirectorio.Definicion // hace referencia a una definición en otro directorio

HPCC SYSTEMS®

Objetivo del tutorial

- Servicio de consulta inmobiliaria:

propertysearchservice-manager-roxie.1 Response

Dataset: Result 1

	lastname	firstname	id	middlename	namesuffix	filedate	gender	birthdate	propcount	personid	propertyid	house number	house number suffix	predir	street	streettype	postdir	apt	city	state	zip
1	TAYLOR	ANJILLETTE	1740051266736247740 5	W		19980619	M	19710614	4	1740051266736247740 5	1965045	830			SKYSAIL	AVE			WARRENSBURG	NY	12885
										1740051266736247740 5	1975231	838		NW	000081ST	AVE		000002	ST LOUIS	MO	63131

HPCC SYSTEMS®

Materia prima

##	id	firstname	lastname	middlename	namesuffix	filedate	gender	birthdate
1	187522928604396	PETRONICA	SPOCK			20030425	F	19290205
2	214582956185891	KIHM	DEMIRTAS	W		19860711	F	19330610
3	345438575926606	DELYNN	MALSCH	T		20000311	M	19700426
4	562092156665191	FOLAKE	KOSTMAN	G		20070922	M	19681006
5	599574955213581	ORA	HUBERT			20111011	M	
6	630037699819979	KUOR	LUHCS			20100402	M	
7	638971319693497	ADEREMI	HOWD			20000422	M	
8	1028541850646460	IRA	DUNHAMPEARS			20130512	F	19861204
9	1096143903819059	TAMASINE	LUECKE	G		20071229	M	
10	1151459511906416	SHARNAE	LITINAS	E		19981017	M	19690104

##	personid	propertyid	house_number	house_number_suffix	predir	street	streettype	postdir	apt	city	state	zip
1	187522928604396	828195	144			MCKIERNAN	DR			WALNUT CREEK	CA	9459
2	187522928604396	1144455	281			CENTER	ST			BALTIMORE	MD	2113
3	187522928604396	1494347	483			NEWTON	RD			FLAGSTAFF	AZ	8601
4	187522928604396	1910847	802			HATCHERY	CT			WOODLAND	WA	9867
5	187522928604396	4267562	5007		E	ROY ROGERS	RD			TROY	MI	4808
6	187522928604396	4888602	7607			PEBBLESTONE	DR		000009	KERNVILLE	CA	9323
7	214582956185891	54135	4			WAINWRIGHT	DR			NORTH FORT MYERS	FL	3391
8	214582956185891	762012	125			SHIPYARD	DR		000150	MELBOURNE VILLAGE	FL	3290
9	214582956185891	2331721	1190			LITTLEOAK	DR			HOUSTON	TX	7701
10	214582956185891	3276109	2506			MEADOW	DR			LA QUINTA	CA	9225

- **People (~279 k)**
- **Property (~1.6 Mi)**
- **Taxdata (~6 Mi)**

##	propertyid	document_year	total_val_calc	land_val_calc	improvement_value_calc	assd_total_val	tax_amount	mkt_total_val	mkt_land_val	mkt_improvement_val
1	1	0000	101400	17600	83800	101400	0	0	0	0
2	1	0000	107600	17600	90000	107600	0	0	0	0
3	3	0000	0	0	0	0	0	0	0	0
4	3	0000	51353	8259	43094	51353	0	0	0	0
5	3	2006	107000	21400	85600	10700	61840	107000	21400	85600
6	4	0000	1852	1852	0	0	2870	1852	1852	0
7	4	0000	1852	1852	0	0	2928	1852	1852	0
8	4	2004	50500	10100	40400	4895	59978	50500	10100	40400
9	4	2004	50500	10100	40400	5050	62154	50500	10100	40400
10	4	2013	89000	17800	71200	8900	75690	89000	17800	71200

HPCC SYSTEMS®

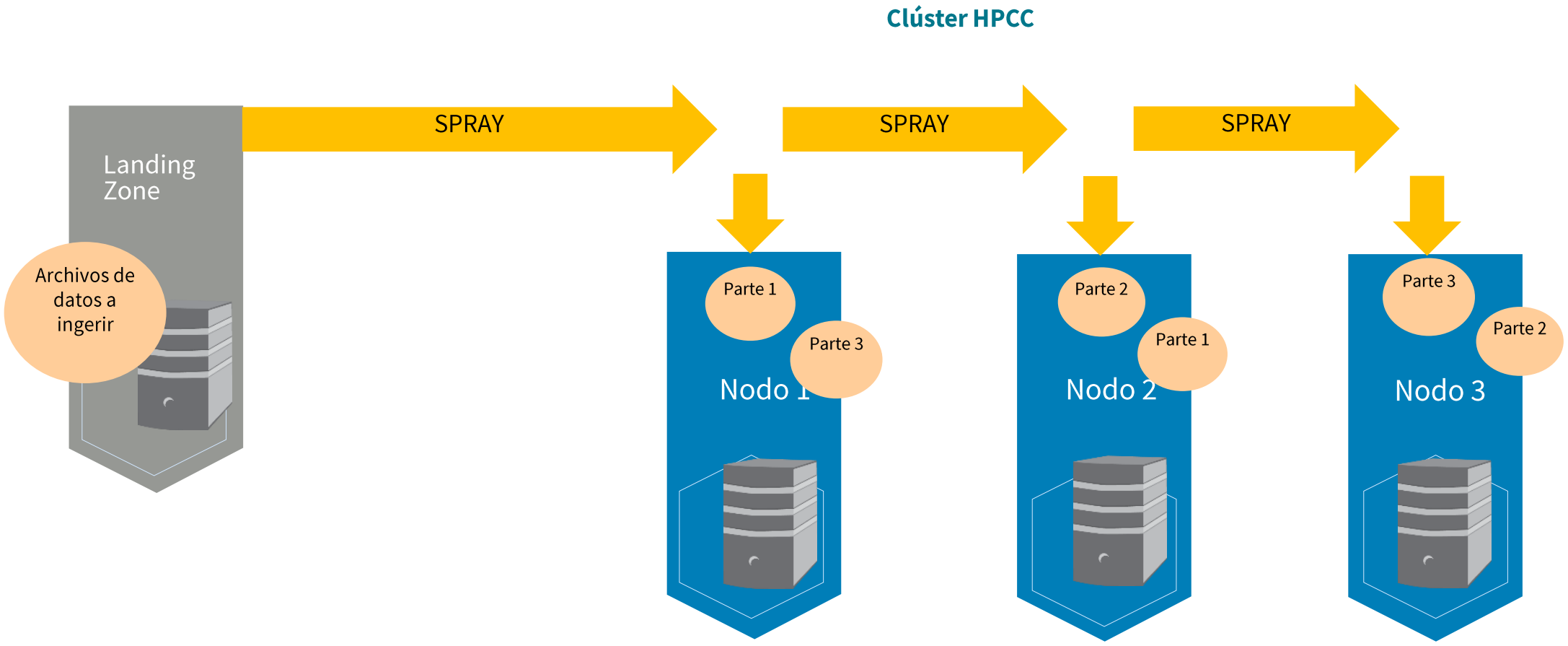
Definición de extracción:

Importar y limpiar datos de diferentes fuentes

- Importación de datos brutos
- Definición de la estructura de datos
- Análisis del perfil de datos

HPCC SYSTEMS®

Extracción: spray de datos

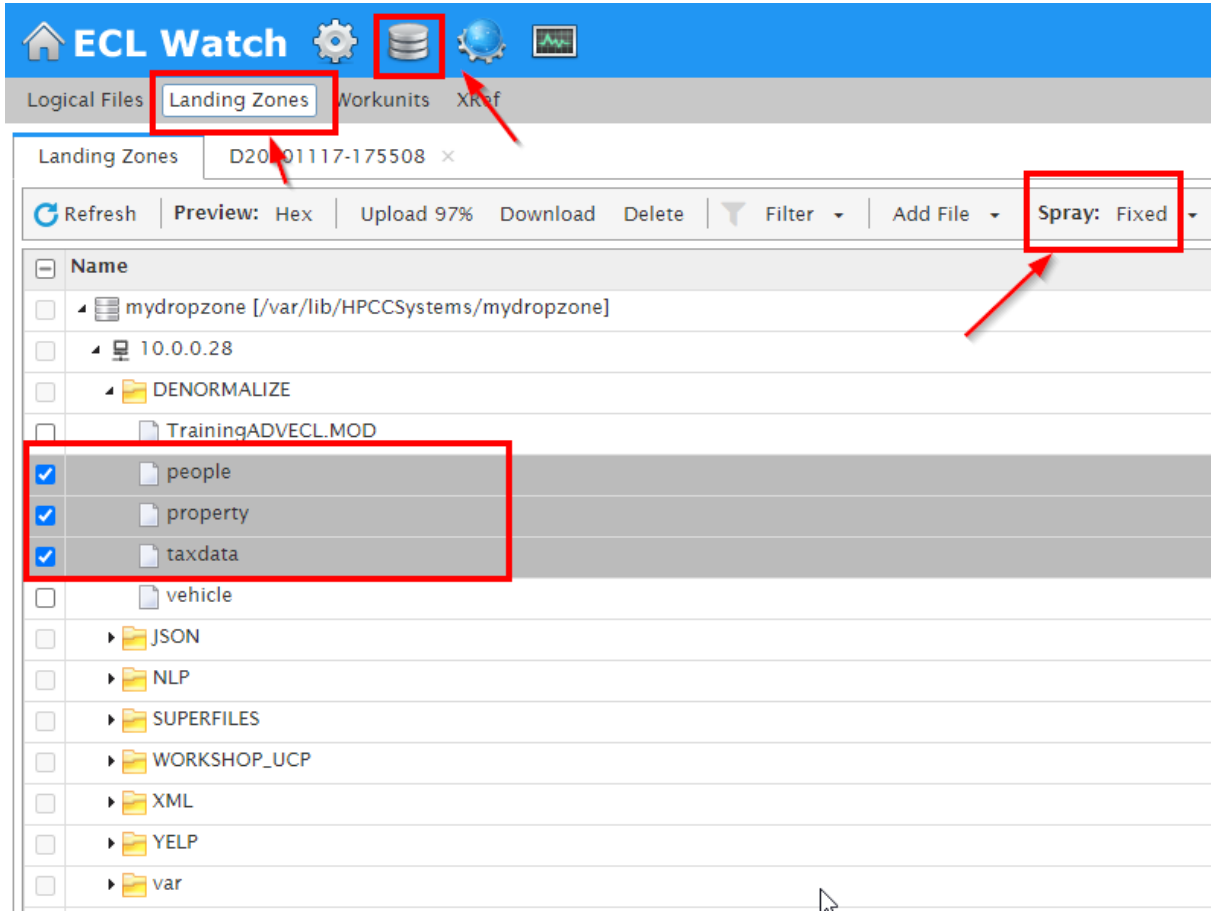


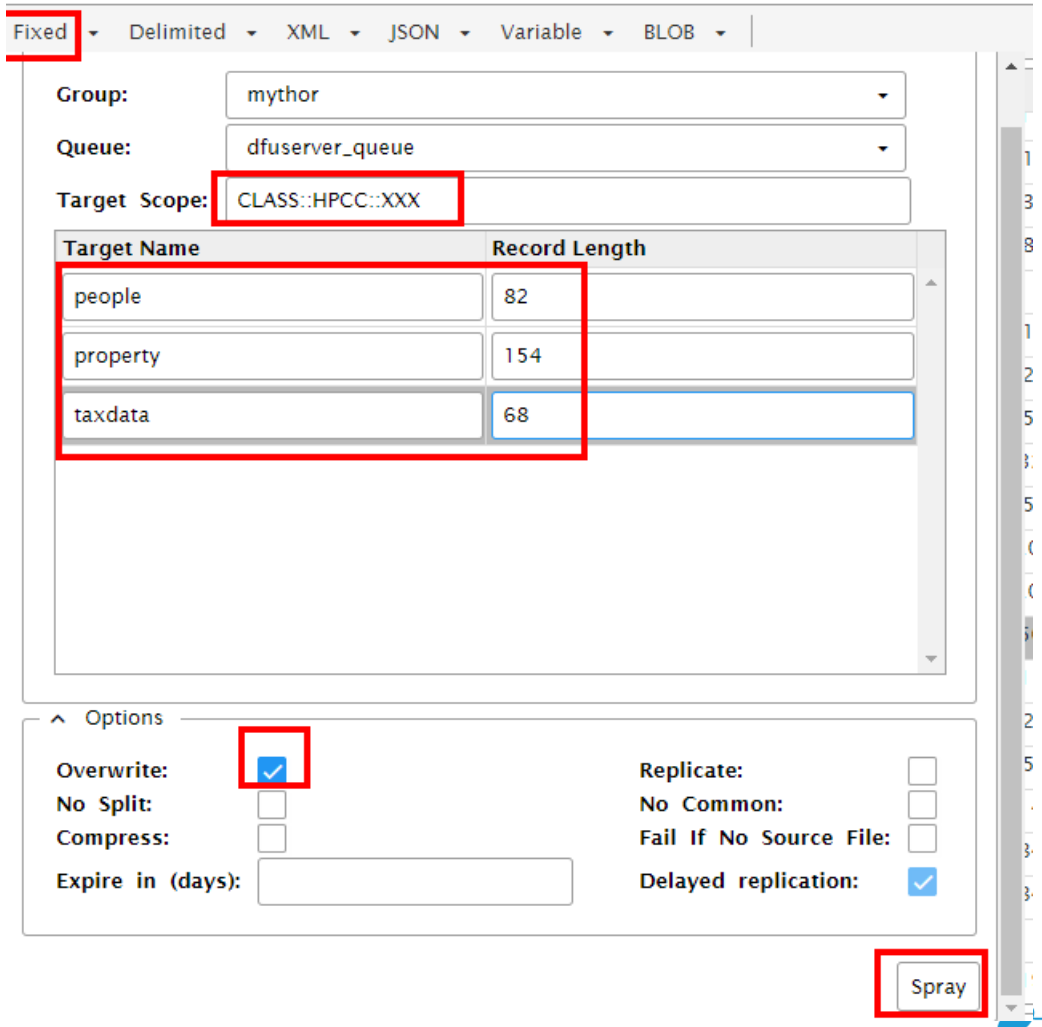
Las partes del archivo se referencian en ECL como un solo archivo lógico ...

HPCC SYSTEMS®

Extracción: Ejecutando el spray

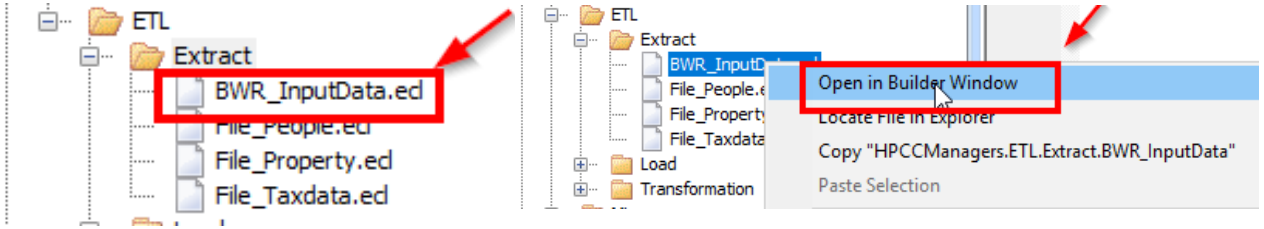
http://3.139.124.33:8010/ (ECL Watch)



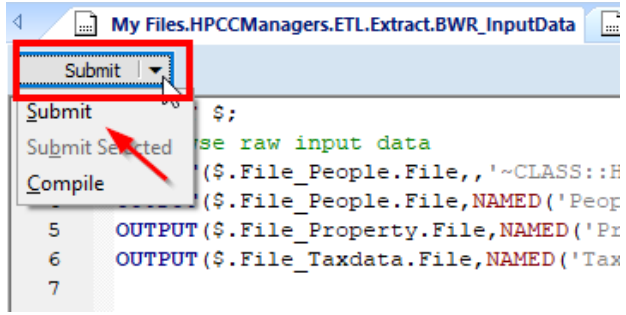


HPCC SYSTEMS®

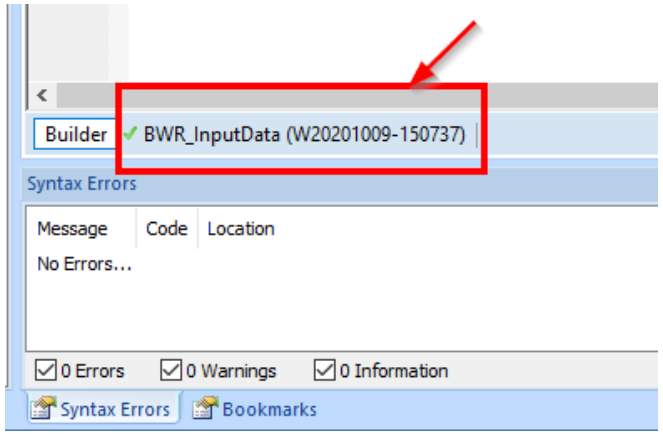
Extracción: comprobación del spray



##	id	firstname	lastname	middlename	namesuffix	filedate	gender	birthdate
1	187522928604396	PETRONICA	SPOCK			20030425	F	19290205
2	214582956185891	KIHM	DEMIRTAS	W		19860711	F	19330610
3	345438575926606	DELYNN	MALSCH	T		20000311	M	19700426
4	562092156665191	FOLAKE	KOSTMAN	G		20070922	M	19681006
5	599574955213581	ORA	HUBERT			20111011	M	
6	630037699819979	KUOR	LUHCS			20100402	M	
7	638971319693497	ADEREMI	HOWD			20000422	M	
8	1028541850646460	IRA	DUNHAMPEARS			20130512	F	19861204
9	1096143903819059	TAMASINE	LUECKE	G		20071229	M	
10	1151459511906416	SHARNAE	LITINAS	E		19981017	M	19690104



##	personid	propertyid	house_number	house_number_suffix	predir	street	streettype	postdir	apt	city	state	zip
1	187522928604396	828195	144			MCKIERNAN	DR			WALNUT CREEK	CA	9459
2	187522928604396	1144455	281			CENTER	ST			BALTIMORE	MD	2113
3	187522928604396	1494347	483			NEWTON	RD			FLAGSTAFF	AZ	8601
4	187522928604396	1910847	802			HATCHERY	CT			WOODLAND	WA	9867
5	187522928604396	4267562	5007		E	ROY ROGERS	RD			TROY	MI	4808
6	187522928604396	4888602	7607			PEBBLESTONE	DR		000009	KERNVILLE	CA	9323
7	214582956185891	54135	4			WAINWRIGHT	DR			NORTH FORT MYERS	FL	3391
8	214582956185891	762012	125			SHIPYARD	DR		000150	MELBOURNE VILLAGE	FL	3290
9	214582956185891	2331721	1190			LITTLEOAK	DR			HOUSTON	TX	7701
10	214582956185891	3276109	2506			MEADOW	DR			LA QUINTA	CA	9225

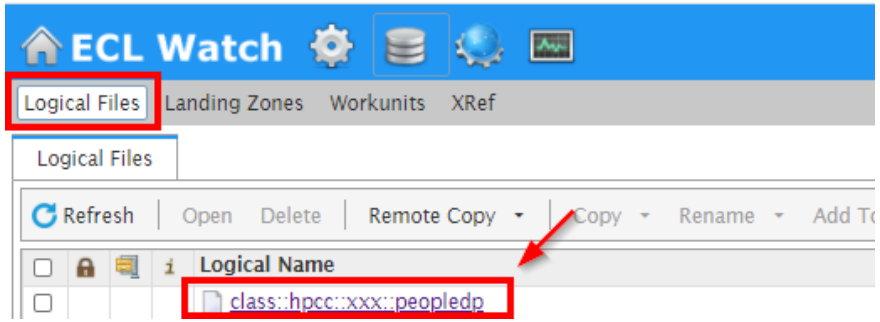


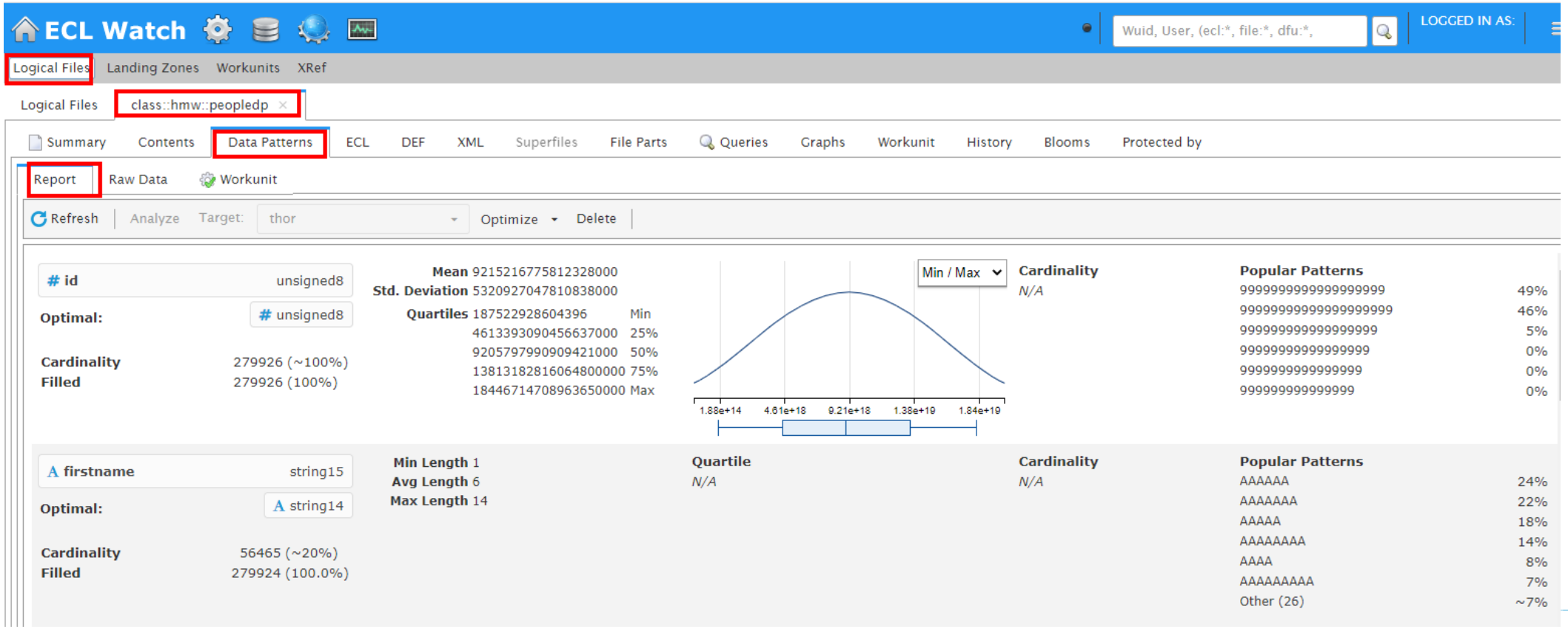
##	propertyid	document_year	total_val_calc	land_val_calc	improvement_value_calc	assd_total_val	tax_amount	mkt_total_val	mkt_land_val	mkt_improvement_val
1	1	0000	101400	17600	83800	101400	0	0	0	0
2	1	0000	107600	17600	90000	107600	0	0	0	0
3	3	0000	0	0	0	0	0	0	0	0
4	3	0000	51353	8259	43094	51353	0	0	0	0
5	3	2006	107000	21400	85600	10700	61840	107000	21400	85600
6	4	0000	1852	1852	0	0	2870	1852	1852	0
7	4	0000	1852	1852	0	0	2928	1852	1852	0
8	4	2004	50500	10100	40400	4895	59978	50500	10100	40400
9	4	2004	50500	10100	40400	5050	62154	50500	10100	40400
10	4	2013	89000	17800	71200	8900	75690	89000	17800	71200

HPCC SYSTEMS®

Extracción: análisis de datos

Reporte de profilling:





HPCC SYSTEMS®

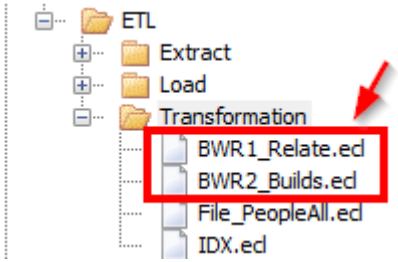
Definición de transformación:

Mapeo y conversión de datos para diseños de registros estandarizados

- Designación de identificadores (recid's)
- Estandarización de campo
- Normalización o desnormalización

HPCC SYSTEMS®

Transformación: desnormalización



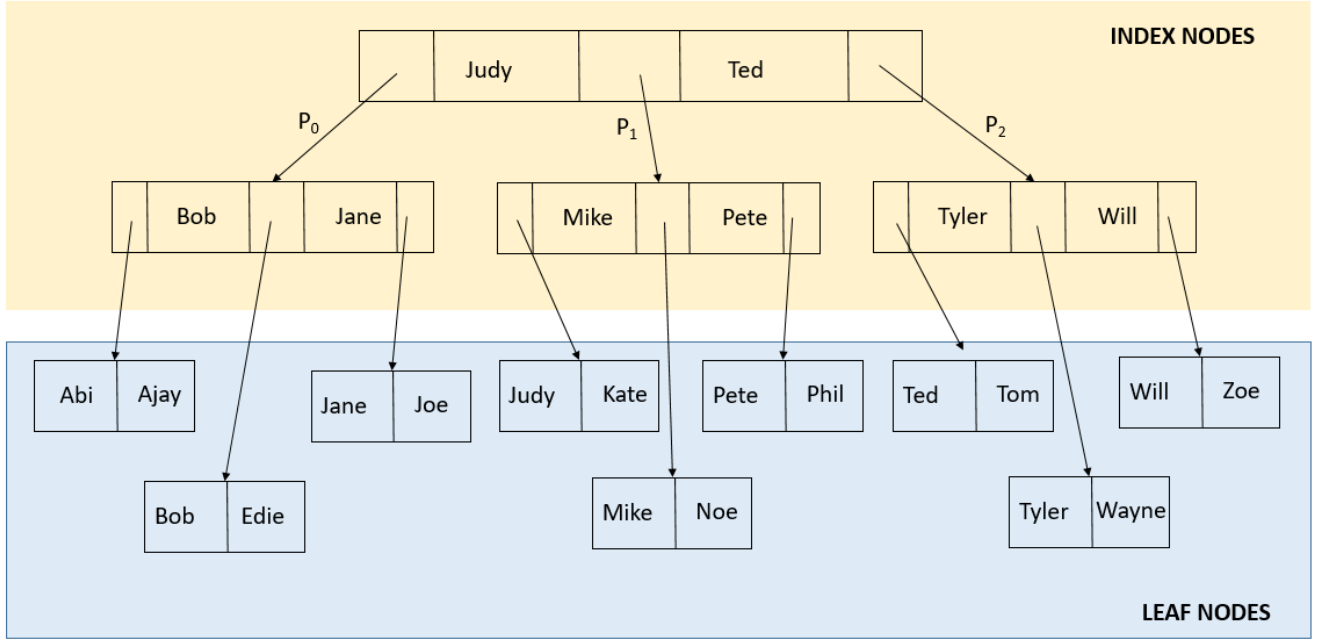
##	id	firstname	lastnar	middlena	namesuffix	filedate	gender	birthdate	propcount	proprecs											
										personid	propertyid	house_number	house_nu	predir	street	streettype	postdir	apt	city	state	zip
1	1875229…	PETRONICA	SPOCK			20030425	F	19290205	6	1875229…	828195	144			MCKIERNAN	DR			WALNUT CREEK	CA	94597
										1875229…	1144455	281			CENTER	ST			BALTIMORE	MD	21136
										1875229…	1494347	483			NEWTON	RD			FLAGSTAFF	AZ	86011

HPCC SYSTEMS®

Definición de carga:

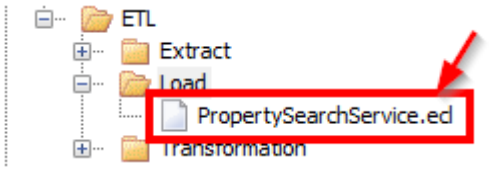
Generación de índices y disponibilidad de datos / consultas.

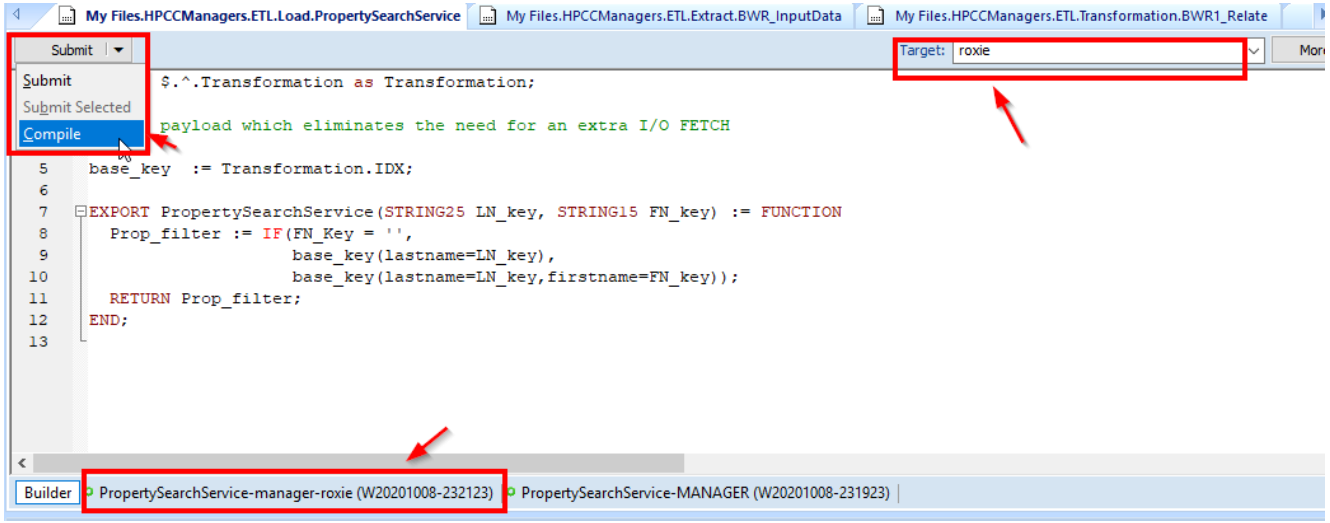
- Crear índices en el clúster THOR
- Disponibilidad de datos y consultas para un clúster ROXIE

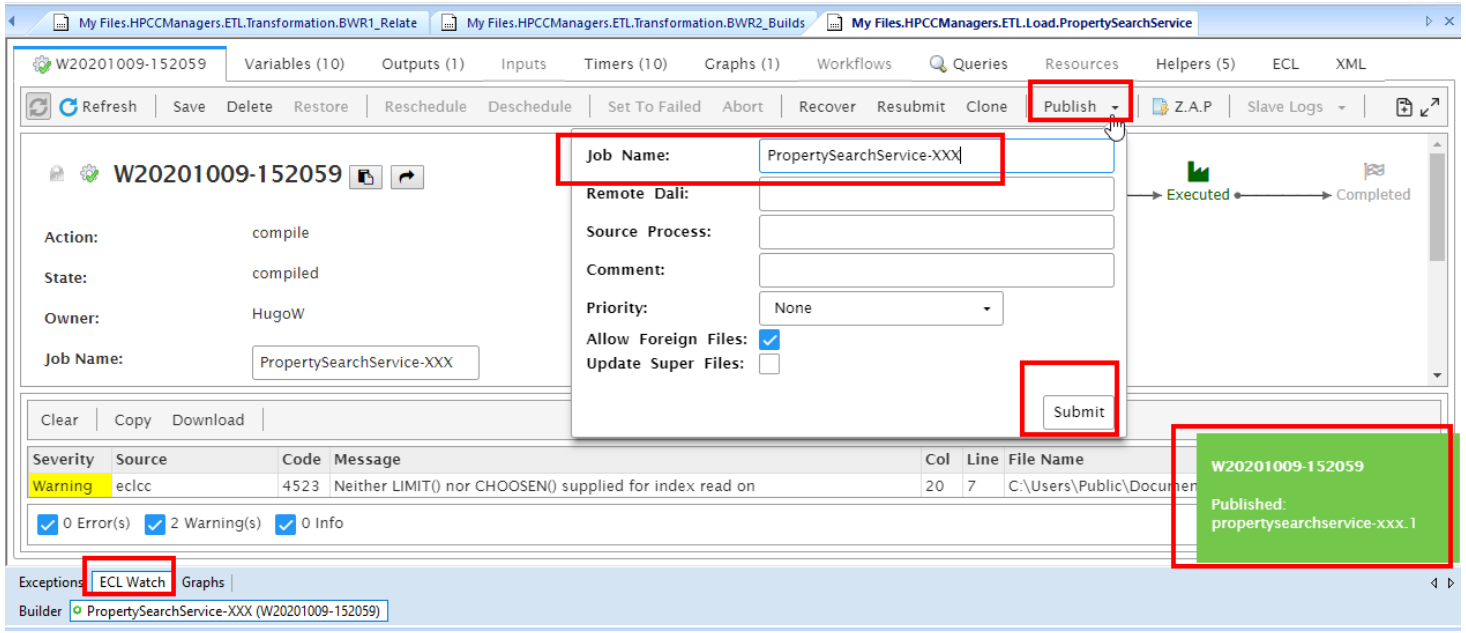


HPCC SYSTEMS®

Cargando: Publicación de la consulta







HPCC SYSTEMS®

¡Servicio disponible para su uso!

ECL Watch

Queries | Package Maps

Queries | propertysearchservice-xxx.1

Refresh | Open | Delete | Suspend | Unsuspend | Activate | Deactivate | Filter | Options | Recreate Query | Published by me

ID	Name	Target
propertysearchservice-xxx.1	propertysearchservice-xxx	roxie

Queries | propertysearchservice-xxx.1

Summary | Errors/Status (1) | Logical Files (1) | Super Files | Libraries Used (0) | Graphs (1) | Resources | Test Pages | W20201009-152059

SOAP | JSON | WSDL | Request Schema | Response Schema | Sample Request | Sample Response | Parameter XML | Legacy Form | Links

Reset

roxie

propertysearchservice-xxx.1 Dynamic Form

PROPERTYSEARCHSERVICE_XXX_1REQUEST

fn_key:

ln_key:

Capture Log Info. Trace Level:

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

HPCC SYSTEMS®

Fin de la parte 2!

HPCC SYSTEMS®

Bundles y aplicaciones

HPCC SYSTEMS®

https://covid19.hpccsystems.com/

HPCC SYSTEMS® OPEN DATABASE LICENSE GITHUB

COMMENTARY/TOP MAP LIST STATS TRENDS HOT SPOTS METRICS TERMS < BACK

THE WORLD

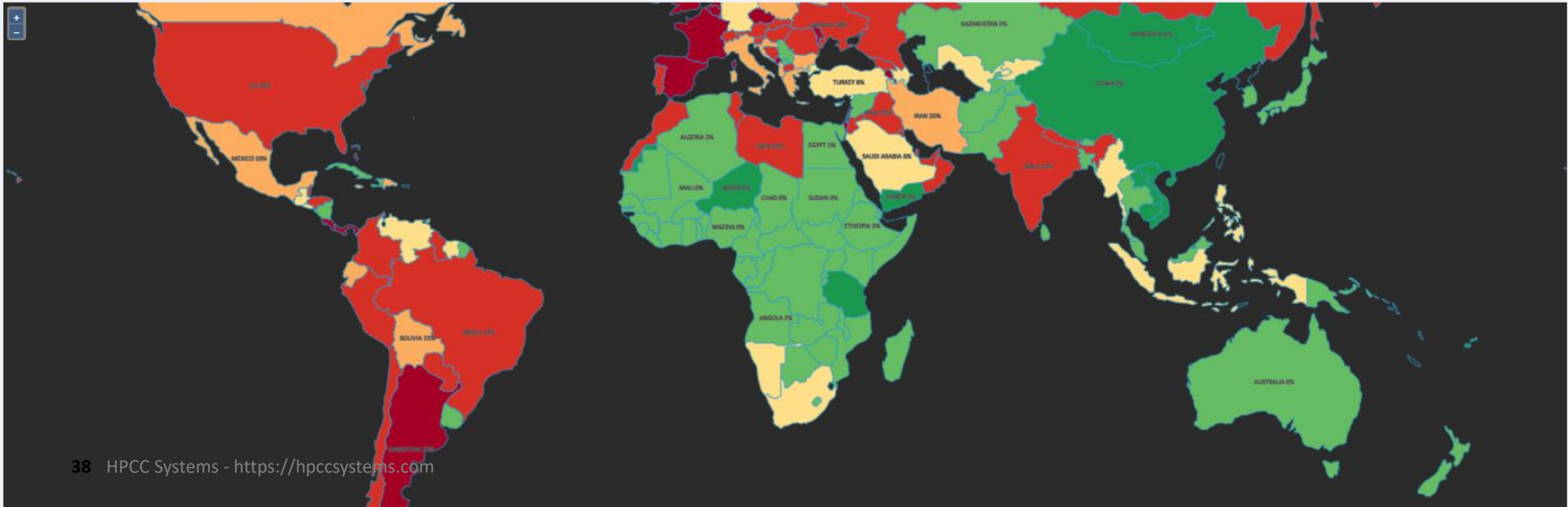
As of Oct 07, 2020, The World remains in a Stabilized state. The infection is steady (R = 1.01). There are currently 3,012,916 active cases. At this rate, expect to see approximately 2,129,361 new cases and 36,551 deaths per week. This is the 3rd surge in infections, which started on the week of Jun 11, 2020. With 2,129,361 new cases, this is the worst week so far for cases during this surge. The Contagion Risk is high at 18.2%. This is the likelihood of meeting an infected person during one hundred random encounters. The Case Fatality Rate (CFR) is estimated as 3.1%. Preliminary estimates suggest that 2% of the population may have been infected and are presumed immune. This is not enough to significantly slow the spread of the virus. This preliminary estimation also implies an Infection Fatality Rate (IFR) of roughly 0.6%. The Short-Term Indicator suggests that the infection is likely to worsen over the course of the next few days.

Maps

Contagion Risk | Infection State | Weekly New Cases | Weekly New Deaths | Cases/100K | Deaths/100K | Cases | Deaths | Legend

Current Historical

Zoom to view more details or click on a location to view details.

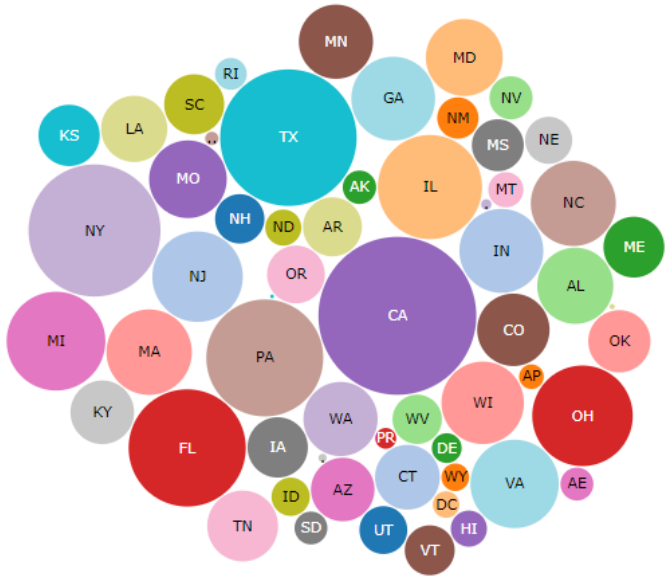


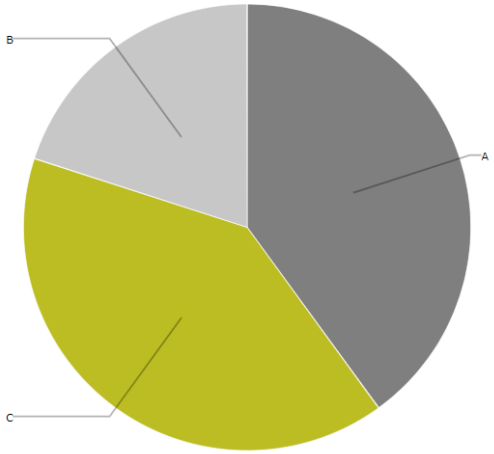
Visualización de datos

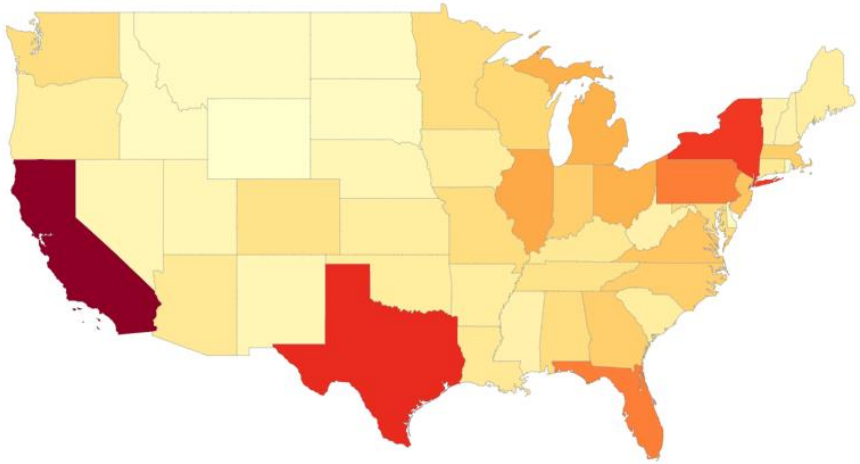
HPCC SYSTEMS®

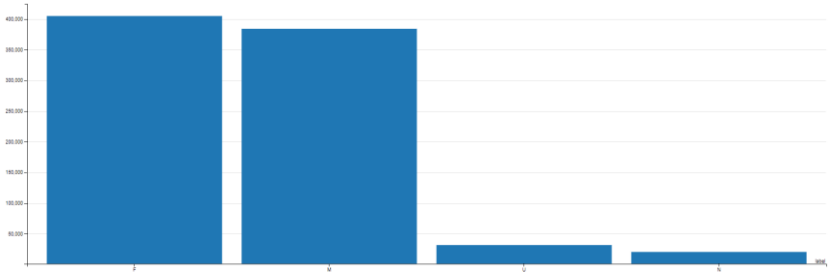
Herramientas de visualización

La plataforma HPCC Systems proporciona herramientas de visualización para datos de salida a través de gráficos y mapas.







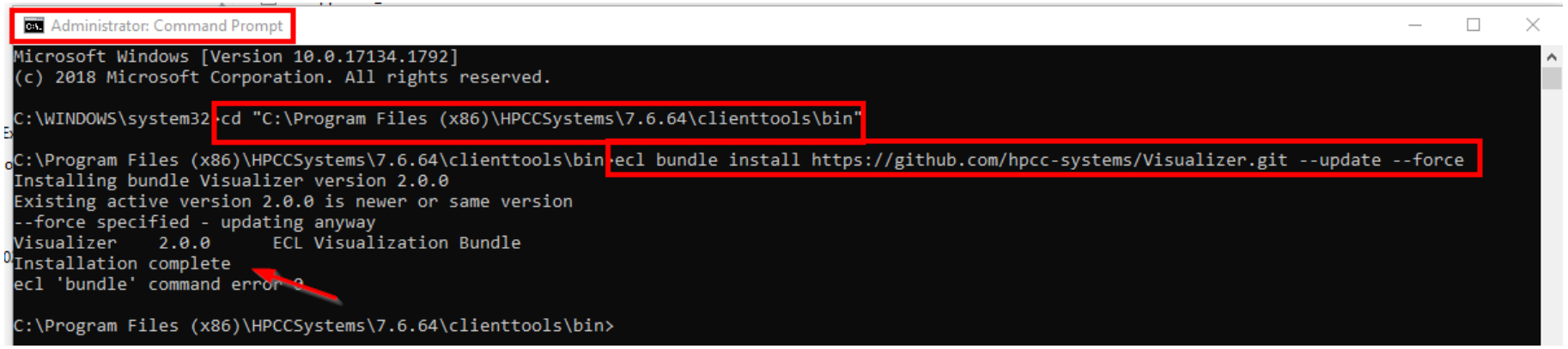


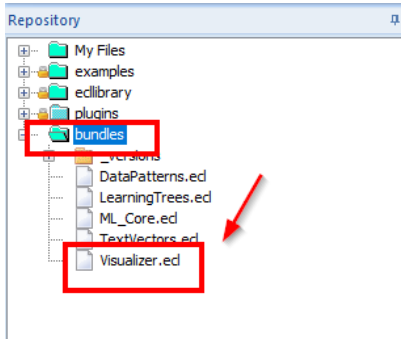
HPCC SYSTEMS®

Herramientas de visualización (cont.)

Los datos se pueden ver utilizando tres métodos:

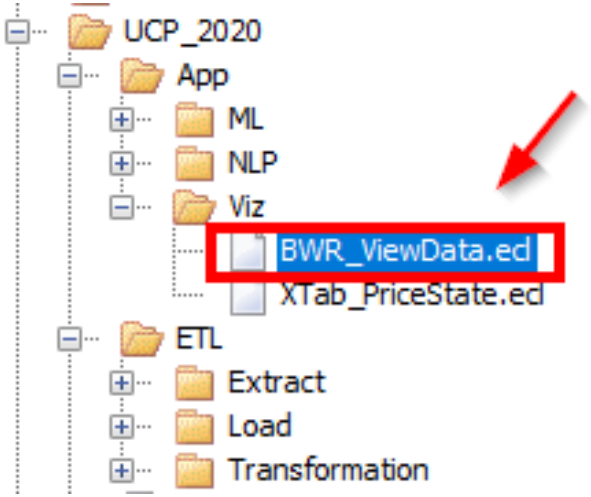
- A través de la herramienta de visualización de Playground.
- A través de la pestaña "Visualizar" en la salida de cualquier unidad de trabajo.
- A través de la pestaña "Recursos" junto con el bundle de visualización ECL.

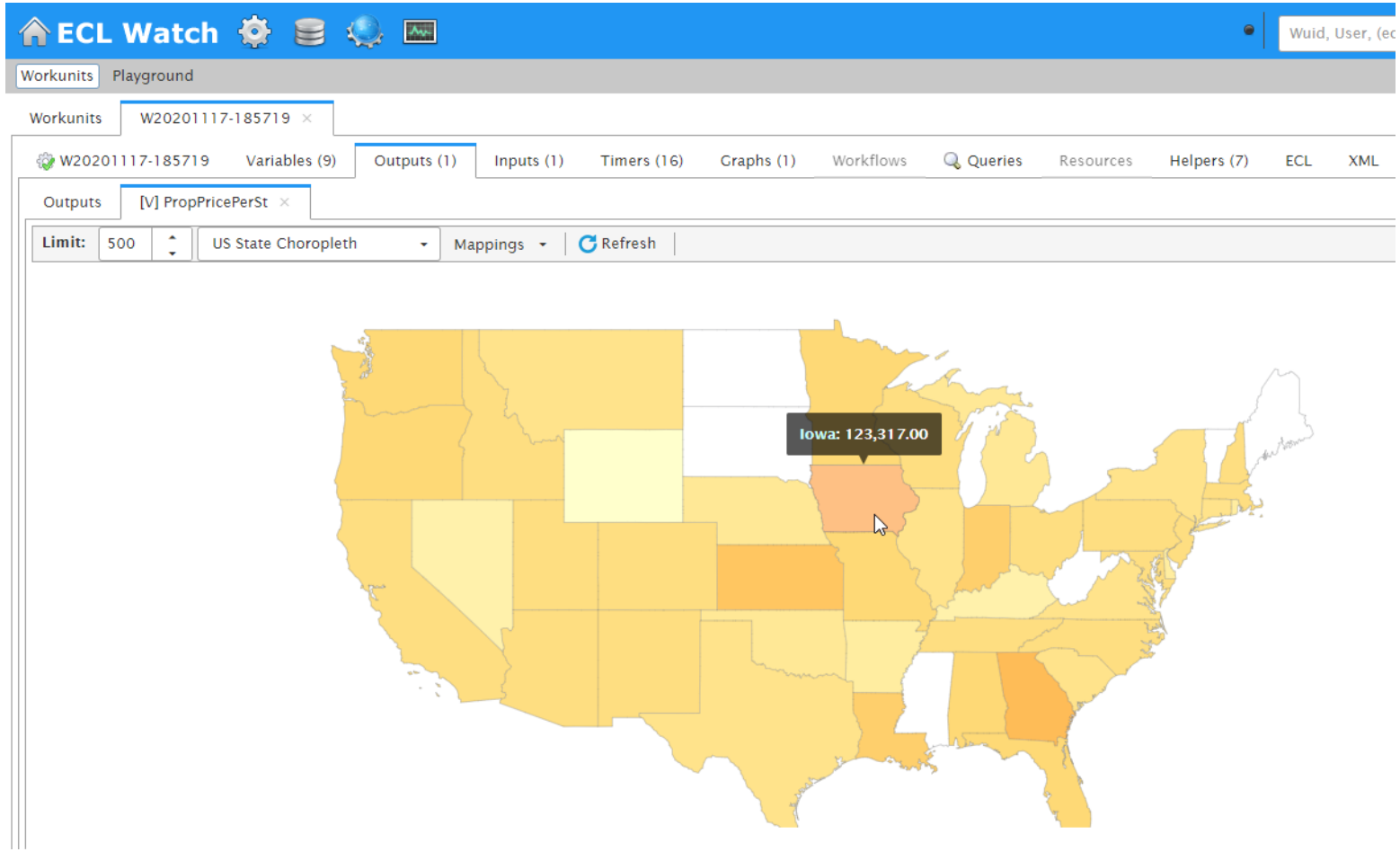




HPCC SYSTEMS®

Los precios medios de propiedades



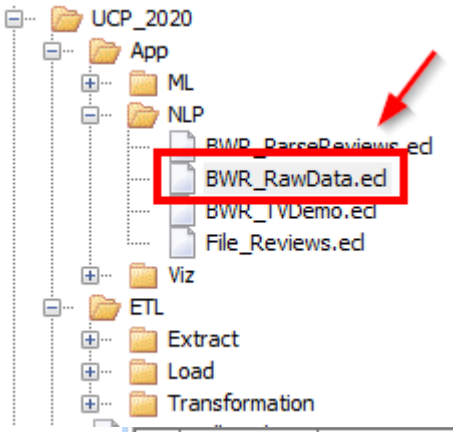


HPCC SYSTEMS®

Procesamiento de Lenguaje Natural (PLN)

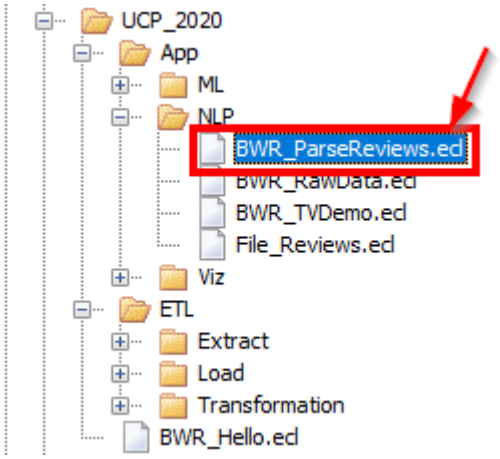
HPCC SYSTEMS®

Los datos brutos (AirBnB)



##	property_id	review_id	review_date	reviewer_id	reviewer_name	review_text
1	7202016	38917982	2015	28943674	Bianca	Cute and cozy place. Perfect location to everything!
2	7202016	39087409	2015	32440555	Frank	Kelly has a great room in a very central location. Beautiful building , architecture an
3	7202016	39820030	2015	37722850	Ian	Very spacious apartment, and in a great neighborhood. This is the kind of apartment I
4	7202016	40813543	2015	33671805	George	Close to Seattle Center and all it has to offer - ballet, theater, museum, Space Needle
5	7202016	41986501	2015	34959538	Ming	Kelly was a great host and very accommodating in a great neighborhood. She has some gre
6	7202016	43979139	2015	1154501	Barent	Kelly was great, place was great, just what I was looking for-clean, simple, well kept
7	7202016	45265631	2015	37853266	Kevin	Kelly was great! Very nice and the neighborhood and place to stay was expected and comf
8	7202016	46749120	2015	24445447	Rick	hola all bnb erz - Just left Seattle where I had a simply fantastic time for the weeken
9	7202016	47783346	2015	249583	Todd	Kelly's place is conveniently located on a quiet street in Lower Queen Anne which is an
10	7202016	48388999	2015	38110731	Tatiana	The place was really nice, clean, and the most important aspect; it was close to everyt
11	7202016	49441269	2015	39852826	Tim	The place was really nice, clean and quiet at night.Clean Linen and Towels were provide
12	7202016	50490194	2015	384855	Tony	The listing was exactly as described! Kelly's place was wonderful and cleen. it was j
13	7202016	53862449	2015	21607838	Jason	Very welcoming and a nicer place to live in the Seattle area

Analizar los comentarios de la propiedades



##	prop_id	subst	verb_prep_adv	adjct
13	7202016	Kelly	was	great
14	7202016	stay	was	expected and comfortable
15	7202016	all	was	good and mega
16	7202016	place	is	conveniently located
17	7202016	schedule	was completely	full and I
18	7202016	place	was really	nice
19	7202016	it	was close to	everything so we
20	7202016	place	was really	nice
21	7202016	Towels	were	provided and the
22	7202016	mattress	was	superb
23	7202016	Neighbourhood	is	practical with a lot of restaurants
24	7202016	Downtown	is	reachable by foot
25	7202016	Kelly	was a	fantastic host

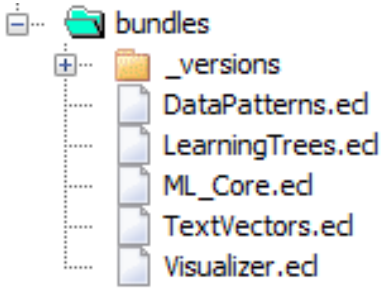
Bundle de Machine Learning

HPCC SYSTEMS®

ML en HPCC Systems

- Bundle validado, compatible y optimizado para la plataforma de sistemas HPCC (https://hpccsystems.com/download/free-modules/machine-learning-library)

- Proceso de instalación:
  - Fácil e independiente de la versión de la plataforma
  - *ecl bundle install https://github.com/hpcc-systems/<nome>.git*



- Soporte de lenguaje en R y Python:
  - Paquetes R
  - Scikit-learn
  - Keras/TensorFlow

HPCC SYSTEMS®

Bundle de ML

- Base:
  - ML_Core: Machine Learning Core (https://github.com/hpcc-systems/ML_Core.git)
  - PBblas: Paralell Block Basic Linear Algebra Subsystem (https://github.com/hpcc-systems/PBblas.git)

- Aprendizaje supervisado
  - LinearRegression: OLS (https://github.com/hpcc-systems/LinearRegression.git)
  - LogisticRegression: binomial/multinomial (https://github.com/hpcc-systems/LogisticRegression.git)
  - GLM: General Linear Model (https://github.com/hpcc-systems/GLM.git)
  - SVM: Support Vector Machines (https://github.com/hpcc-systems/SupportVectorMachines.git)
  - LearningTrees: Árboles de decisión (https://github.com/hpcc-systems/LearningTrees.git)

HPCC SYSTEMS®

Bundle de ML (cont.)

- Aprendizado não-supervisionado
  - K-Means: agrupación de Big Data (https://github.com/hpcc-systems/KMeans.git)
  - DBSCAN: Scalable Paralell Density-Based Spatial Clustering of Applications with Noise (https://github.com/hpcc-systems/dbscan.git)
  - TextVectors: Vectorización de palabras, frases y oraciones (https://github.com/hpcc-systems/TextVectors.git)

- Aprendizaje profundo
  - GNN: Generalized Neural Network (https://github.com/hpcc-systems/GNN.git)

HPCC SYSTEMS®

Soporte a pipeline de ML

1. Definición del problema
2. Extracción de datos
3. Preparación de datos
4. Segregación de datos
5. Entrenamiento de modelos
6. Evaluación y selección de modelos
7. Implementación del modelo

HPCC SYSTEMS®

Ejemplo: Modelo de regressión

“"Dado un conjunto de atributos de una propiedad (ubicación, metraje, año de construcción), ¿cómo predecir su valor?"

propertyid	house_numbe	house_nu	predir	street	streett	postdir	apt	city	state	zip	total_value	assessed_value	year_acquired	land_square_foot	living_square_fee	bedrooms	full_baths
828195	144			MCKIERNAN	DR			WALNUT CREEK	CA	94597	62614	62614	2006	20418	2485	3	2
1144455	281			CENTER	ST			BALTIMORE	MD	21136	105500	10550	2007	4807	1368	0	0
1494347	483			NEWTON	RD			FLAGSTAFF	AZ	86011	2220	2220	0	5654	1011	3	1
1910847	802			HATCHERY	CT			WOODLAND	WA	98674	356000	356000	0	6094	0	2	1
4267562	5007		E	ROY ROGERS	RD			TROY	MI	48085	327253	327253	2007	3484	0	3	0
4888602	7607			PEBBLESTONE	DR		000009	KERNVILLE	CA	93238	732179	732179	2010	19597	6132	6	6
48725	4			LONG	AVE			SUNRISE	FL	33323	271000	271000	2008	6880	2392	4	2
83528	6			TRILLUM	LN			WAYLAND	MA	02193	79889	79889	2007	7657	1657	4	1
94604	7			PARMENTER	AVE			PLYMOUTH	MN	55441	23800	23800	2005	19994	1754	3	2
220326	17			TIMBER	RD			LOS ANGELES	CA	90063	89000	89000	2008	7840	954	3	1
994609	212			FREYER	DR	NE		PHILOMONT	VA	20131	59800	59800	2009	11199	1241	3	0
1836173	724			EASTER	ST			ALLENTOWN	PA	18102	191600	191600	0	9100	2534	4	2
2910797	1903			SADDLE BROOK	DR			CLIO	CA	96106	61610	61610	2007	0	0	0	0
3083959	2158			RIVERSIDE	DR			UPPER MORELA…	PA	19006	90300	0	0	0	1235	3	2
3952189	4040			GRAND VIEW	BLVD		000054	RIO LINDA	CA	95673	0	0	0	2700720	0	0	0
4186238	4726			LAS PALMAS	CT			WAELDER	TX	78959	18816	18816	2009	2159	1320	0	0
4597143	6213			WILSON	RD			ZOLFO SPRINGS	FL	33890	72600	0	0	8496	0	3	1
4624905	6321			STONEWALL	LN			PATERSON	NJ	07514	139880	139880	2008	10454	1391	4	2
92326	7			KNOLLCREST	DR			NARANJA	FL	33032	76214	76214	2008	4800	930	2	0
1792852	704			ERIN	DR			TRABUCO	CA	92678	28010	28010	2007	5200	0	3	1
1843977	728		S	ARLINGTON HE…	RD			BLOOMING GRO…	TX	76626	130400	130400	2007	36154	1629	3	1
4214872	4821			MYRTLE OAK	DR		000025	SAN BERNARDI	CA	92376	[illegible]	0	2007	93654	0	0	0

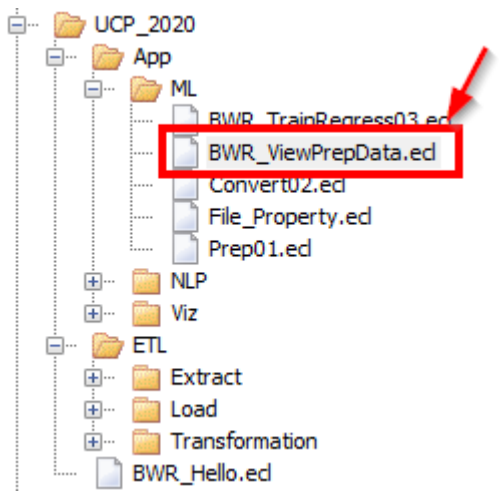
HPCC SYSTEMS®

Instalación del bundle

```
Command Prompt
(c) 2018 Microsoft Corporation. All rights reserved.

C:\Users\watahu01>cd "C:\Program Files (x86)\HPCCSystems\7.6.64\clienttools\bin"

C:\Program Files (x86)\HPCCSystems\7.6.64\clienttools\bin>ecl bundle install https://github.com/hpcc-systems/ML_Core.git --update --force
Installing bundle ML_Core version 3.2.2
Existing active version 3.2.2 is newer or same version
--force specified - updating anyway
ML_Core       3.2.2      Common definitions for Machine Learning
Installation complete
ecl 'bundle' command error 0

C:\Program Files (x86)\HPCCSystems\7.6.64\clienttools\bin>ecl bundle install https://github.com/hpcc-systems/LearningTrees.git --update --force
Installing bundle LearningTrees version 1.1.1
Existing active version 1.1.1 is newer or same version
--force specified - updating anyway
LearningTrees 1.1.1      LearningTrees Bundle for Tree-based Machine Learning
Installation complete
ecl 'bundle' command error 0
```


HPCC SYSTEMS®

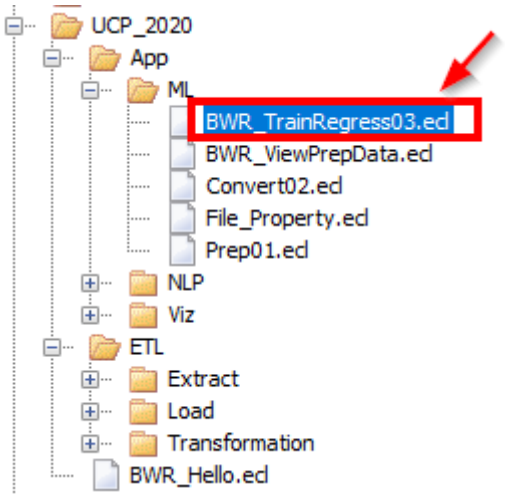
Los datos de la regressión



##	propertyid	zip	assessed_value	year_acquired	land_square_footage	living_square_feet	bedrooms	full_baths	half_baths	year_built	total_value	rnd
1	828195	94597	62614	2011	20418	2485	3	2	1	2009	62614	2681399375
2	4888602	93238	732179	2015	19597	6132	6	6	0	2010	732179	2016489933
3	762012	32904	96300	2015	18000	2357	4	2	1	2006	96300	1951014038
4	1565512	93300	245000	2012	5649	1149	3	2	0	1967	245000	2234351926
5	1837309	49333	9470	2015	154202	2284	3	2	0	1986	32672	3900215832
6	4542536	13159	148550	2015	7890	3113	4	2	1	2011	148550	1662479476
7	1892631	76136	840	2012	392040	2004	3	0	0	2003	188810	1533850982
8	3541423	95375	30300	2015	6534	2795	2	2	0	2002	331100	1750945454
9	3831369	15234	109163	2015	7143	768	2	1	0	1962	109163	3161440663
10	978550	83610	223900	2011	16552	2470	3	2	1	1999	223900	2377635704
11	4836980	7670	107500	2013	15681	1582	3	2	0	1987	107500	1313131942
12	769894	66062	119835	2013	6000	1447	3	2	0	1982	119835	3266895727
13	900426	93420	196639	2014	8700	2504	3	2	1	2008	196639	1223955664
14	1636831	23060	216240	2013	6540	1178	3	2	0	1977	216240	510070662
15	689976	78023	38440	2012	4420	1509	3	2	1	1982	104590	1855145411
16	3827323	20695	93600	2013	4000	930	3	1	1	1995	93600	2094279832
17	4401557	93637	25271	2011	217800	1916	3	2	1	2008	259447	2346788527
18	1502153	12250	154675	2016	1240	1105	3	2	1	2001	154675	2741141577
19	1829065	95403	44590	2011	16727	1896	4	2	1	2005	127400	937358568
20	1981287	34949	202953	2016	7700	1452	3	2	0	1977	202953	796231433

HPCC SYSTEMS®

La calidad del modelo



My Files.HPCCManagers.ML.BWR_TrainRegress03

Submit

```
IMPORT LearningTrees AS LT;
IMPORT ML_Core;
IMPORT $;

myLearnerR2     := LT.RegressionForest(,,,[1]);
myModelR2       := myLearnerR2.GetModel($.Convert02.myIndTrainDataNF, $.Convert02.myDepTrainDataNF);

predictedDeps2 := myLearnerR2.Predict(myModelR2, $.Convert02.myIndTestDataNF);
OUTPUT(predictedDeps2);


assessmentR2    := ML_Core.Analysis.Regression.Accuracy(predictedDeps2, $.Convert02.myDepTestDataNF);
OUTPUT(assessmentR2);
```

My Files.HPCCManagers.ML.BWR_TrainRegress03

##	wi	regressor	r2	mse	rmse
1	1	1	0.864343702015207	2534988213.899637	50348.66645602083

HPCC SYSTEMS®

Fin de la parte 3!

HPCC SYSTEMS®

Resumen del taller

- ✓ Descripción general de la plataforma HPCC Systems
  - ✓ Definición
  - ✓ Histórico
  - ✓ Componentes

- ✓ Familiaridad con el proceso ETL en HPCC
  - ✓ Extracción
  - ✓ Transformación
  - ✓ Carga

- ✓ Comprender las aplicaciones
  - ✓ Visualización
  - ✓ PLN
  - ✓ ML

HPCC SYSTEMS®

Entrenamiento online: learn.lexisnexis.com/hpcc

- Introducción a ECL (parte 1)
  - Conceptos y consultas
- Introducción a ECL (parte 2)
  - ETL con ECL
- ECL avanzado (parte 1)
  - Datos relacionales
- ECL avanzado (parte 2)
  - Superarchivos, XML / JSON y PLN
- ECL aplicado
  - Generación y automatización de código ECL
- ROXIE ECL (parte 1)
  - Índices y consultas
- ROXIE ECL (parte 2)
  - Optimización de consultas
- Machine Learning con HPCC Systems
  - Fundamentación para uso de los bundles
- Administración de sistemas
  - Conceptos básicos y funcionamiento
- HPCC para gerentes
  - Descripción general y aplicaciones de la plataforma

HPCC SYSTEMS®

Enlaces útiles

- Sitio principal : **hpccsystems.com/es**
- Primeros pasos : **hpccsystems.com/es/about**
- Download: **hpccsystems.com/es/download**
- Foro de la Comunidad: **hpccsystems.com/forums**

Unete a la communidad

Regístrese en hpccsystems.com

HPCC SYSTEMS®