

A large iceberg floats in the ocean under a clear blue sky. The iceberg has a prominent, jagged peak on the left side and a long, flat base extending to the right. The water is a deep blue, and the sky is a lighter blue.

**90% of your Big Data Problem, isn't Big Data.
It's the ability to handle Big Data for better insight.**

HPCC Systems Machine Learning

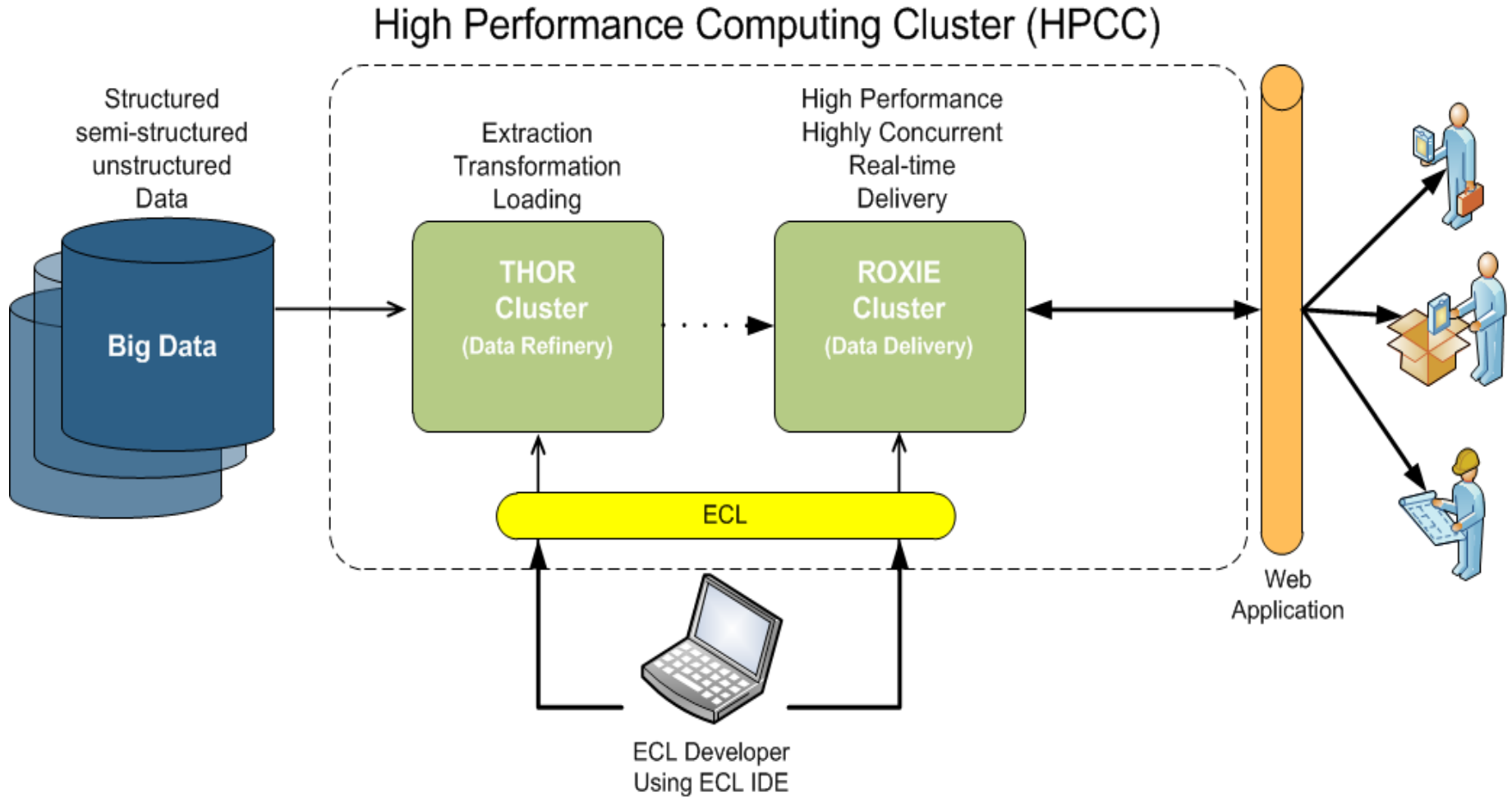
**Edin Muharemagic, Ph.D.
Architect and Data Scientist
HPCC Systems**

LexisNexis Risk Solutions

- More than **15 years** of Big Data experience
- Provides **information** solutions to enterprise customers
- Generates about \$1.4 billion in revenue
- Has been using the HPCC Systems platform (High-Performance Computing Cluster) for over 10 years

HPCC Systems

- Launched in June 2011 (http://hpccsystems.com/about-us/press_center/lexisnexis-announces-hpcc-systems)
- Open source, and enterprise-proven distributed Big Data analytics platform
- To help enterprises manage Big Data at every step in the Complete Big Data Value Chain



Consistent and elegant HW&SW architecture across the complete platform:
<http://hpccsystems.com/Why-HPCC/How-it-works>

Thor – Data Refinery Cluster

- Distributed parallel processing, Distributed File System, Scales from 1-1000s
- Optimized for Extraction, Transformation, Loading, Sorting, Indexing

Roxie – Query Cluster

- Distributed parallel processing, Distributed File System, Scales from 1-1000s
- Optimized for concurrent query processing

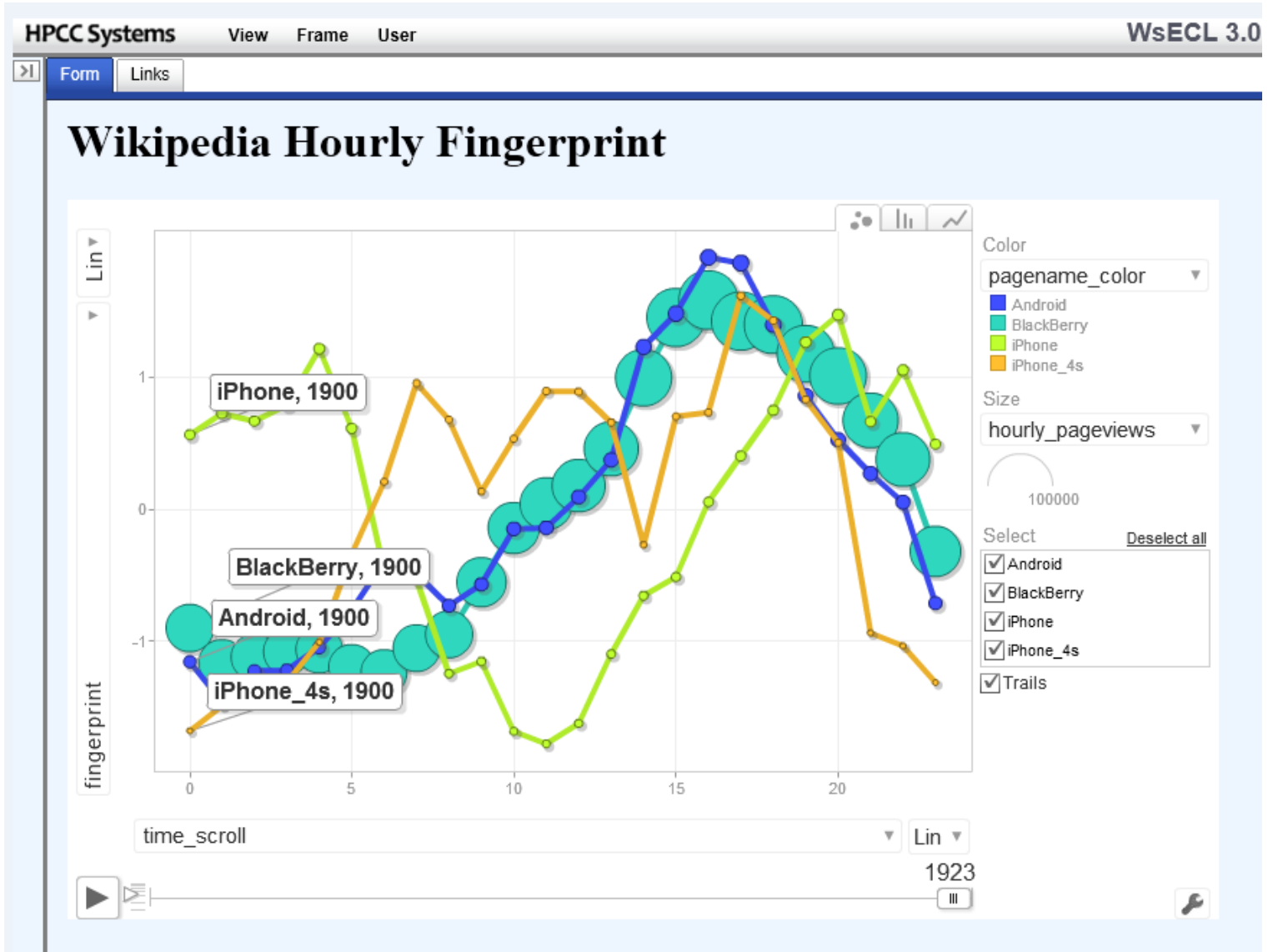
ECL – Enterprise Control Language

- Transparent and Implicitly parallel programming language
- Non-procedural and dataflow oriented
- Modular, reusable, extensible
- Built-in PARSE and PATTERN ops suitable for Natural Language Processing

ECL IDE

- Modern IDE used to code, debug, and monitor ECL programs

- A fully distributed and extensible set of Machine Learning techniques for Big Data
- State of the art algorithms in each of the Machine Learning domains, including supervised and unsupervised learning:
 - Correlation: Covariance, Pearson, Kendall
 - Classifiers: Naïve Bayes, Perceptron, Logistic
 - Clustering: Kmeans, Agglomerative (Hierarchical)
 - Regression: Ordinary Least Squares, Polynomial
 - Document manipulation: Tokenize, N-gram extraction, NLP
 - Associations: AprioriN, EclatN
 - Distribution: Uniform, Normal, Poisson, Exponential, Binomial,
 - Discretize
 - Field Aggregates
- Distributed and parallel underlying linear algebra library
 - Matrix Factorization: SVD, Eig, Lanczos, PCA, Cholesky, Householder, LU



- If you don't already use the HPCC platform and/or ECL IDE and the Client Tools, you must download and install them before downloading the ML libraries
 - Download and install the relevant HPCC platform for your needs: (<http://hpccsystems.com/download/free-community-edition>)
 - Download and install the ECL IDE and Client Tools (<http://hpccsystems.com/download/free-community-edition/ecl-ide-and-client-tools>)
 - Take a look at the ECL programmers guide and ECL language reference guides, <http://hpccsystems.com/community/docs/learning-ecl>.
 - Take a look at tutorials designed to get you started using data on the HPCC Systems, <http://hpccsystems.com/community/docs/tutorials>.
 - Go to the Machine Learning page of the HPCC Systems website, <http://hpccsystems.com/ml> and click on Download and Get Started.