

Global Big Data Conference

BIG DATA BOOTCAMP

December 9th, 10th & 11th 2016

Tampa



Tampa Convention Center, 333 S Franklin St, Tampa, FL 33602

www.globalbigdataconference.com

Twitter : @bigdataconf

IRI, The CoSort Company

Vendor Background

- ISV specializing in data management and data protection
- Known since 1978 for “big data” transformation speed
- 7 of 8 software products share 1 metadata and Eclipse GUI
- A ‘top big data provider’ (CIO Review & Insight Success)
- Headquartered 1 hour southeast of Orlando, FL
- Resellers in more than 40 international cities
- Customers in every industry with big and/or sensitive data

Global Big Data Conference

Selected IRI Customers

IRI customers process and protect data off the mainframe, for DW ETL/ODS ops, and in PII protection (privacy law compliance) initiatives. Hadoop use is optional. Most work with big and/or sensitive financial, call/click, or healthcare data.

BMO Bank of Montreal

RBS
The Royal Bank of Scotland

WesBank citi

Bank of America



KEB Hana Bank

CredibanCo

Fidelity INVESTMENTS

VISA

SUNGARD

EQUIFAX

Experian

EPSILON

Prudential

Aflac

BlueCross BlueShield

Allianz

THE HARTFORD

HSBC

LG

SAMSUNG

American Airlines

החברה לאוטומציה
במנהל השלטון המקומי



GENERAL GROUP

Hewlett Packard Enterprise



vivo

Mercedes-Benz

UNIVERSAL

Comcast

NTTグループ

MEDICK digital

The Walt Disney Company

accenture
High performance. Delivered.
1-800 flowers.com

Global Big Data Conference

IRI Data Manager Suite

FACT
IRI FAsT extraCT

Speed DB unloads for archival, migration, reorg and ETL

- Extract tables to flat files in parallel using SQL queries
- Convert and re-format to change data types and layouts
- Create the data definitions for IRI software and DB loads
- Pipe to CoSort and DB loaders for faster reorg and ETL

CoSORT
THE OPEN SYSTEMS STANDARD

Speed or replace batch, BI, ETL, sort, and SQL programs

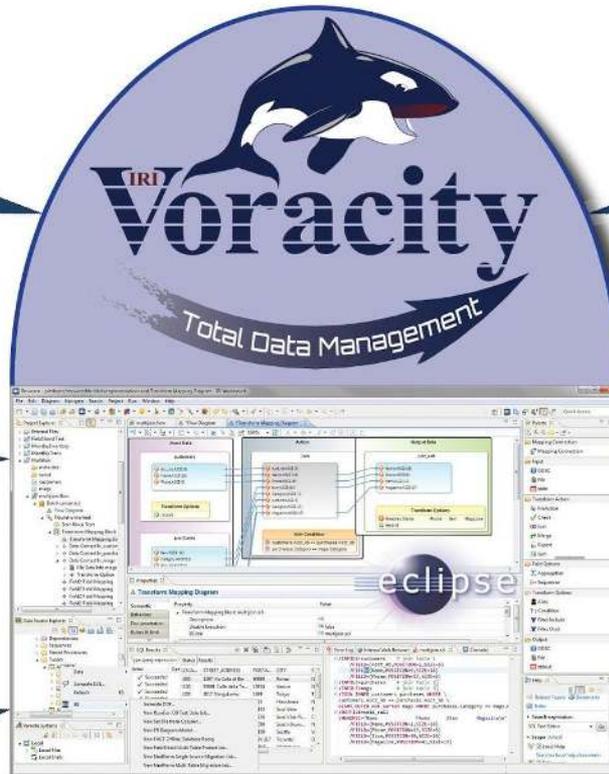
- Filter, sort, join, aggregate, pivot, cleanse, lookup, calc, etc.
- Map, migrate, federate and replicate data from 125 sources
- Segment data, capture changes, report details / summaries
- Analyze changing dimensions, support complex transforms

NextForm
Data & Database Migration

Unlock data and move between apps, DBs, and platforms

- Convert, federate, remap, and replicate legacy data
- Migrate data between databases and create new tables
- Segment file formats, data types, and endian conditions
- Search and structure data in "dark data" documents

Embedded or callable analytics:
BIRT, JupiterOne, NextCoder, R



Consolidate tools and tasks to process, protect, prototype, present

- Discover, define, and govern data in legacy and new sources
- Combine data integration, migration, protection, and analytics
- Exploit CoSort and Hadoop engines for optimum throughput
- Leverage Eclipse familiarity, functionality, and extensibility

IRI
The CoSort Company

IRI Data Protector Suite

RowGen
Safe Intelligent Test Data

Prototype DBs and ETL, stress-test, outsource, benchmark

- Use real data models and formats, not production data
- Combine generation and selection, create new formats
- Preserve referential integrity and frequency distributions
- Feed test DBs, files, and custom reports simultaneously

FieldShield
Data Masking & Encryption Solutions

Comply with privacy laws, nullify breaches, govern data

- Select shields for each field per business rules
- De-ID, encrypt, hash, mask, pseudonym, random, token
- Apply cross-table rules to save time and referential integrity
- Create an XML audit log of each job to verify compliance

CellShield
Data Masking Add-In for Excel

Profile and protect PAN/PHI/PII in Excel spreadsheets

- Search and save patterns to discover sensitive data
- Locate, report, and open all found ranges in the LAN
- Click to encrypt, mask or pseudonymize data directly
- Auto-log protections to verify privacy law compliance

Chakra Max
Smart Data-Centric Audit & Protection

Define, monitor, block, and audit DB access

- High-volume, data-centric audit and protection (DCAP)
- Monitor, block, alert, and log users in real-time
- Low-impact on DB performance and availability
- Classify and dynamically mask sensitive data with RBAC

Global Big Data Conference

Address the Challenges of Big Data

Volume

BI and analytic tools choke on high volumes; they drag, hang or crash

*Voracity blends and prepares data for analytic tools via **fast, combinatory transforms** like: filter, sort, join, aggregate and segment. Programs built on the CoSort SortCL language hand off digestible data chunks or cubes to BIRT, Qlik, R, SAS, Splunk, Tableau, etc.*

Variety

The myriad of structured and unstructured sources is beyond most tools

*Voracity either natively, or through partner drivers, connects to and integrates **>125 data sources** on premise or in the cloud. They can be structured, semi-structured, or unstructured, and static and streaming.*

Velocity

IOT logs, dark data, CDRs, etc. are generated too fast for analysis

*Voracity processes **streaming data** from: web services and brokers (MQTT, Kafka); pipes; in Hadoop Spark or Storm; SQL; and, through memory via input procedure calls to CoSort. Voracity's built-in task launcher can also run jobs in near-real-time.*

Veracity

Garbage in=garbage out: low quality data jeopardizes analytic value

*Voracity's data **discovery and quality** features let you: search for strings and patterns, do fuzzy matching, validate, scrub, enrich, and unify data for DW/BI, MDM, and analytics.*

Value

Without tackling the above, you won't get analytic value from big data

*Voracity runs with or without Hadoop on commodity hardware under an **affordable subscription** model based only on the number (not size) of servers. Its **Eclipse GUI** is free, familiar, and flexible, to speed learning and time-to-solution.*

Global Big Data Conference

Supported Data Sources/Targets:

Amazon EMR Hive	FinancialForce	Marketo	Pivotal Greenplum
Apache Cassandra	Force.com apps	MongoDB	Pivotal HD Hive
Apache Hadoop Hive	Hortonworks Hive	MS Dynamics CRM	Salesforce.com
Cloudera CDH Hive	Hubspot	MS SQL Azure	ServiceMAX
Cloudera Impala	Lightning Connect	Oracle Eloqua	Spark SQL
Database.com	MapR Hive	Oracle Service Cloud	Veeva CRM

... plus 'legacy list' on next 2 pages >>

Global Big Data Conference

Acucobol Vision	Delimited	MaxDB	SQL Server
Altibase (FACT)	Derby (WB)	Mongo (WB)	SQLite
ASN.1 TAP3	ESDS	MF-ISAM	Sybase ASA/E & IQ
BIRT DB (WB)	Excel (WB)	WF Var. Length	Tibero (WB)
BIRT Hive (WB)	ELF web logs	MySQL	Teradata (WB)
BIRT JDBC (WB)	Fixed	Oracle	Text
BIRT POJO (WB)	Heap / print	Outlook (WB)	UTF-8 & 16
C-ISAM	HSQLDB (WB)	PDF (WB)	Variable Block
CLF web logs	IDX 3, 4 & 8	PostgreSQL	Variable Sequential
CSV	Informix	Powerpoint (WB)	VSAM MVS (UniKix)
DB2 (UDB)	Ingres	Record Sequential	Web Services (WB)
DB2 for i5/OS (WB)	LDIF	RTF (WB)	Word (WB)
DB2 for z/OS (WB)	Line Sequential	SQL Anywhere	XML

Global Big Data Conference

Access	D3	GA-Power 95, R91	K-ISAM	Pathway	RMS
Adabas	Datacom	Gemstone	Knowledgeman	PDS	Reality/X
Advanced Pick	Dataflex	GENESIS	KSDS	PervasiveSQL	RRDS
ALLBASE	Db4o	Gigabase	Lotus	Pick/Pick64+	SAP HANA
Alpha5	dBase	H2	Manman	PI-Open	Sequoia
Amazon RDS	Desktop Adapter	IDMS	Mentor / pro	Powerflex	Sharebase
Azure	DL/1	IDS	MO	Powerhouse	Supra
BizTalk	DSM	Image	Model 204	Progress	Terracotta
Cache	Enscribe	IMS	Mumps	QueryObject	Total
Clipper	Enterprise Adapter	Interbase	MyBase	rBase	Ultimate
Codasyl	FileMaker	Intersystems	Netezza	R83	UltPlus
CorVision	Firebird	ISM	NonStop SQL	Rdb	Unidata
ConceptBase	Focus	Jasmine	ObjectStore	REALITY	Universe
D-ISAM	FoxPro	JBase	Paradox	Red Brick	VSAM VSE

Global Big Data Conference

Sources

- Big Data**
- Call Detail Records**
 ASN.1 Formats
- Cloud & SaaS**
- Databases**
- Files & Pipes**
 COBOL, CSV, LDIF, LS-RS-VS, MFVL, Text, VB, Vision, XML
- Mainframe**
 Adabas, Datacom, IDMS, IMS, ISAM, Pick, Unidata, VSAM, etc.
- Semi & Unstructured**
- Other Sources**
 Custom Apps, ETL/ELT Tools, Packaged Apps, Web Logs

DISCOVER

Data Classification
 Dark Data Search
 DB & File Profiling
 ER Diagramming
 Metadata Definitions
 Metadata Forensics
 Multi-Method Search

Targets

- Big Data**
- BI & Analytic Tools**
- Cloud & SaaS**
- Custom Reports**
 Detail & summary reports
- Databases**
- Files & Pipes**
 COBOL, CSV, LDIF, LS-RS-VS, MFVL, Text, VB, Vision, XML
- Other Targets**
 Custom Apps, Data & SpreadMarts, ETL/ELT Tools, Federated Views, Packaged Apps, Test Suites

INTEGRATE

Public/Private Mashup
 Change Data Capture
 Bulk DB Un/Load
 Data Federation
 One Pass ETL

MIGRATE

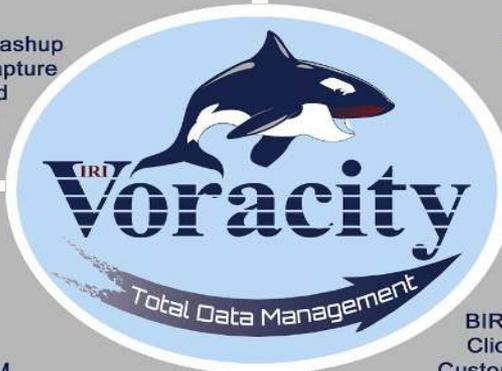
Data & File Types
 Endianness
 Databases
 ETL Jobs
 JCL Sorts

GOVERN

Data Lineage
 Data Masking
 Data Quality
 Metadata & MDM
 Test Data Generation

ANALYZE

Embedded BI
 BIRT & Splunk Feeds
 Clickstream Analytics
 Customer Segmentation
 Slowly Changing Dimensions



DESIGN

ADS Mapping Manager
 Form Editors
 Graphical Dialogs
 Outlines & Palettes
 Script Editors
 Visual Workflow
 Wizards & Rules

DEPLOY

CoSort CLI/API (SMP)
 Eclipse & Other Job Launchers
 Java, Paques, SQL
 MapReduce (Grid)
 Spark (In-Memory)
 Storm (Streaming)
 Tez (Batch)



Global Big Data Conference



DISCOVER

INTEGRATE

MIGRATE

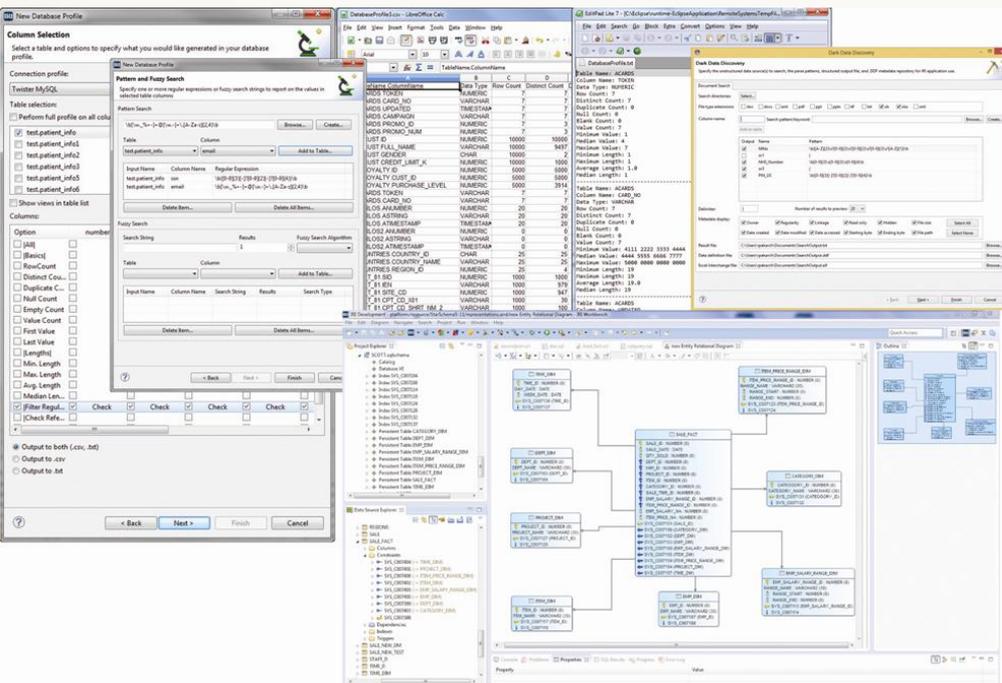
GOVERN

ANALYZE



Voracity includes PII discovery facilities for multi-source data **classification**, string (literal or in-dictionary), pattern, and fuzzy-match **searches**, statistical **reports**, and automatic **metadata** creation. Fit-for-purpose wizards in Voracity perform:

- Data classification, with rule matcher libraries
- DB profiling and E-R diagramming
- Dark data discovery and structuring, with forensic metadata display
- Flat-file statistical and value searching
- Metadata discovery and definition
- Metadata sharing, lineage tracking, etc.



Global Big Data Conference



DISCOVER

INTEGRATE

MIGRATE

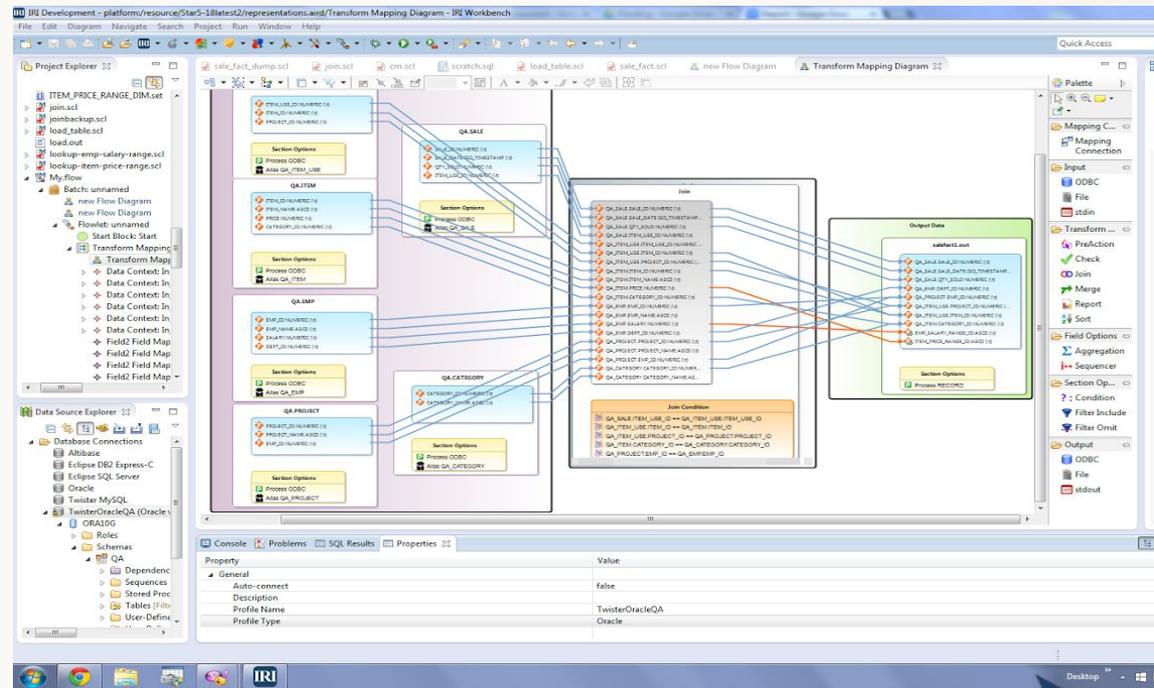
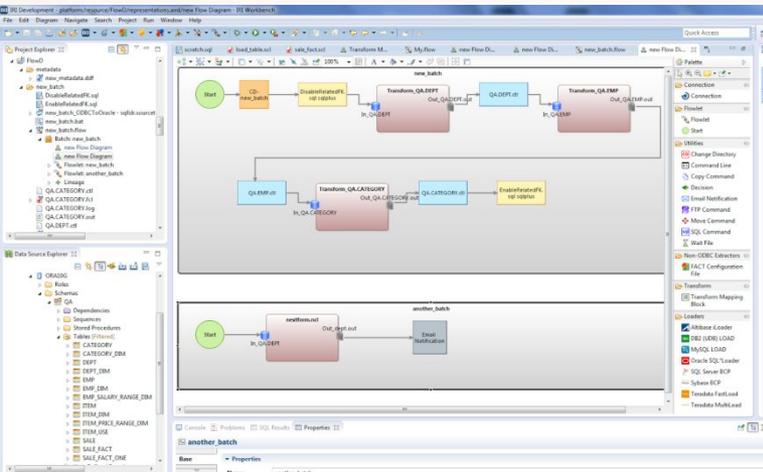
GOVERN

ANALYZE



Voracity combines fast ETL engines and task consolidation techniques with simple metadata in Eclipse that's shared by all IRI software and other products, like AnalytiX DS for ETL code conversion. You can use Voracity to speed or *re-platform* megavendor tools, and optimize:

- EDW, LDW, ODS, data lakes
- Data quality (cleansing)
- VLDB unload/reorg/load jobs
- SCD, CDC, pivoting, unification



Global Big Data Conference



DISCOVER

INTEGRATE

MIGRATE

GOVERN

ANALYZE



Job Design ...

In addition to GUI wizards, diagrams, and dialogs, you can also hand-code the underlying 4GL programs in Voracity's syntax-aware editor.

This job sorts and filters an employee CSV file into two target files, while also redacting ID #'s and commissions, and encrypting the salary.

```
hadoop-demo - IRI Development - demo-hdfs/employee-dept.scl - IRI Workbench
File Edit Navigate Search Project Run Window Help
employee-dept.scl dept50_out dept80_out
# Generated with IRI Workbench - New Sort Job
#
# Author: claudiai
# Created: 2016-11-29 11:37:07
#
@/INFILE=Employees.data|
/PROCESS=DELIMITED
/ALIAS=Employees
/FIELD=(EMPLOYEE_ID, TYPE=ASCII, POSITION=1, SEPARATOR=",")
/FIELD=(FIRST_NAME, TYPE=ASCII, POSITION=2, SEPARATOR=",")
/FIELD=(LAST_NAME, TYPE=ASCII, POSITION=3, SEPARATOR=",")
/FIELD=(EMAIL, TYPE=ASCII, POSITION=4, SEPARATOR=",")
/FIELD=(PHONE_NUMBER, TYPE=ASCII, POSITION=5, SEPARATOR=",")
/FIELD=(HIRE_DATE, TYPE=ASCII, POSITION=6, SEPARATOR=",")
/FIELD=(JOB_ID, TYPE=ASCII, POSITION=7, SEPARATOR=",")
/FIELD=(SALARY, TYPE=ASCII, POSITION=8, SEPARATOR=",")
/FIELD=(COMMISSION_PCT, TYPE=ASCII, POSITION=9, SEPARATOR=",")
/FIELD=(MANAGER_ID, TYPE=ASCII, POSITION=10, SEPARATOR=",")
/FIELD=(DEPARTMENT_ID, TYPE=ASCII, POSITION=11, SEPARATOR=",")

/SORT
/KEY=(LAST_NAME, TYPE=ASCII)

@/OUTFILE=dept50_out
/PROCESS=DELIMITED
/INCLUDE WHERE DEPARTMENT_ID EQ 50
/FIELD=(MASK_EMPLOYEE_ID=replace_chars(EMPLOYEE_ID, "*", 1, 5), TYPE=ASCII, POSITION=1, SEPARATOR="\t")
/FIELD=(FIRST_NAME, TYPE=ASCII, POSITION=2, SEPARATOR="\t")
/FIELD=(LAST_NAME, TYPE=ASCII, POSITION=3, SEPARATOR=" ")
/FIELD=(ENC_FP_SALARY=enc_fp_aes256_alphanum(SALARY, "secret"), TYPE=ASCII, POSITION=4, SEPARATOR="\t")

@/OUTFILE=dept80_out
/PROCESS=DELIMITED
/INCLUDE WHERE DEPARTMENT_ID EQ 80
/FIELD=(MASK_EMPLOYEE_ID=replace_chars(EMPLOYEE_ID, "*", 1, 5), TYPE=ASCII, POSITION=1, SEPARATOR=" ")
/FIELD=(FIRST_NAME, TYPE=ASCII, POSITION=2, SEPARATOR=" ")
/FIELD=(LAST_NAME, TYPE=ASCII, POSITION=3, SEPARATOR=" ")
/FIELD=(ENC_FP_SALARY=enc_fp_aes256_alphanum(SALARY, "secret"), TYPE=ASCII, POSITION=4, SEPARATOR=" ")
/FIELD=(MASK_COMMISSION_PCT=replace_chars(COMMISSION_PCT, "*", 3, 1), TYPE=ASCII, POSITION=5, SEPARATOR=" ")
Writable Insert 7:23
```

Global Big Data Conference



DISCOVER

INTEGRATE

MIGRATE

GOVERN

ANALYZE



Job Deployment ...

Voracity's 4GL scripts run on the command line or in batch from the GUI or shell.

BIRT or Splunk can also run them as they report or index.

Voracity can also schedule and run them seamlessly in MR2, Spark, Spark Stream, Storm or Tez.

The screenshot displays the IRI Workbench interface. At the top, a window titled 'Transform Mapping Diagram' shows a data flow from 'personalInformation2' through a 'Sort' action to two output datasets: 'female_personal_info_encrypted' and 'male_personal_info_encrypted'. The 'Sort' action is configured with 'personal_info.NAME.ASCHI (0)' as the key. The 'female...' output is filtered by 'GENDER EQ 0', and the 'male...' output is filtered by 'GENDER EQ 1'. Below this, a 'Run Configurations' dialog is open for a configuration named 'Hadoop_demo'. It shows the file 'Hadoop/HadoopDemo.scl' and the working directory '/user/java/demo/'. The 'Engines' section has 'Map Reduce 2' selected. A green arrow points from the text 'Map once, deploy anywhere' to the working directory field. At the bottom, two 'Data Viewer' windows show the output data. The left viewer shows the 'male_personal_info_encrypted' data with columns for ID, name, and address. The right viewer shows the 'female_personal_info_encrypted' data with columns for ID, name, and address.

Global Big Data Conference



DISCOVER

INTEGRATE

MIGRATE

GOVERN

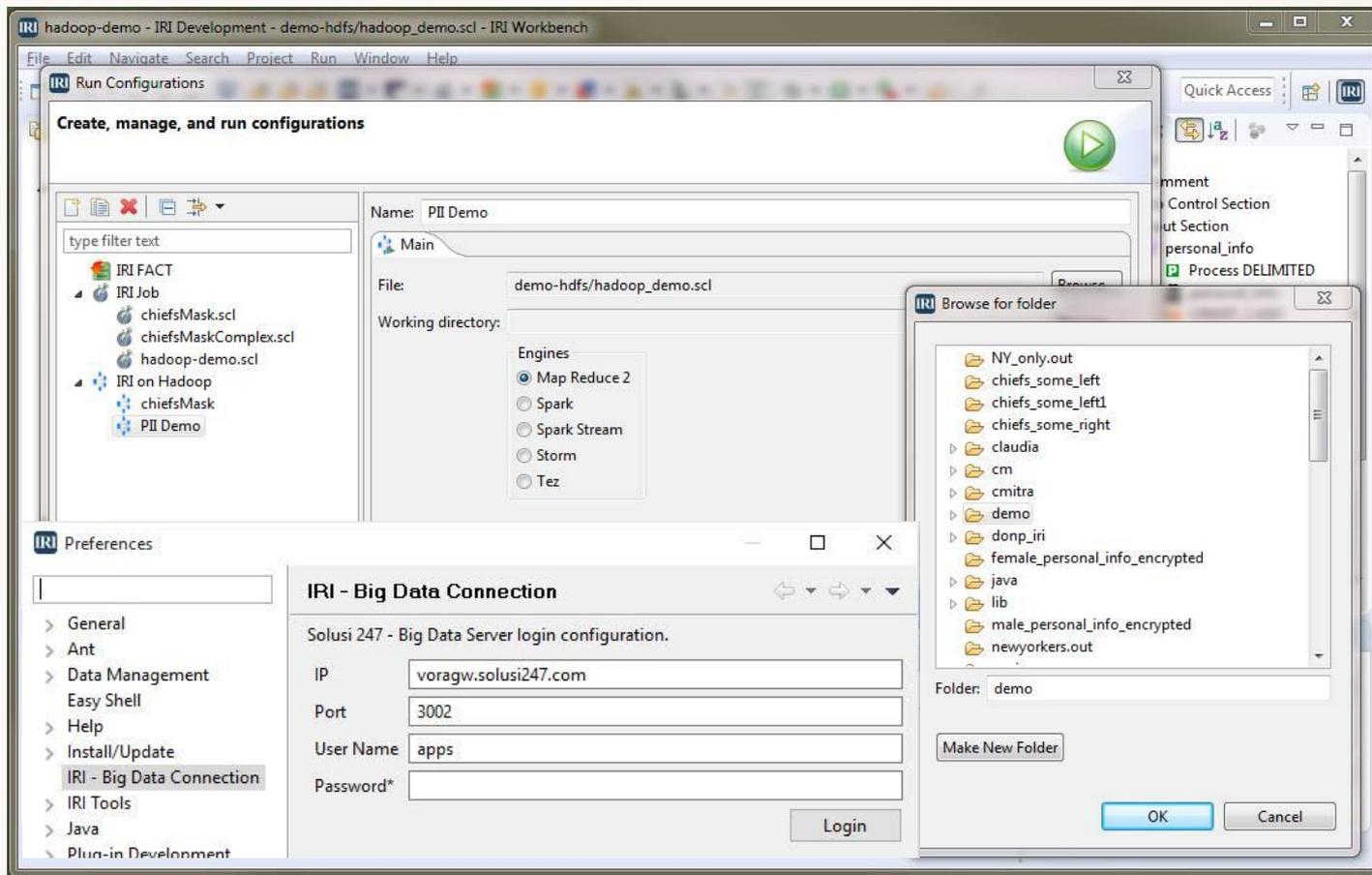
ANALYZE



Preparing a run configuration for Hadoop ...

Once our gateway is open, we can tell any job to run in Hadoop.

Here, we specify MR2 as the engine, and our working directory in HDFS.



Global Big Data Conference



DISCOVER

INTEGRATE

MIGRATE

GOVERN

ANALYZE



The Job Manager view shows our Hadoop job running, plus the status of other jobs.

ID	Name	Engine	Status	User	Start	End
0000005-161130113927475-oozie-oozi-W	demo	MR2	RUNNING	yava	Thu, 01 Dec 2016 19:58:18 GMT	null
0000004-161130113927475-oozie-oozi-W	demo	MR2	SUCCEEDED	yava	Thu, 01 Dec 2016 19:43:28 GMT	Thu, 01 Dec 2016 19:44:38 GMT
0000003-161130113927475-oozie-oozi-W	demo	MR2	SUCCEEDED	yava	Thu, 01 Dec 2016 19:39:14 GMT	Thu, 01 Dec 2016 19:40:24 GMT
0000002-161130113927475-oozie-oozi-W	demo	MR2	SUCCEEDED	yava	Thu, 01 Dec 2016 19:30:12 GMT	Thu, 01 Dec 2016 19:31:22 GMT
0000001-161130113927475-oozie-oozi-W	demo	MR2	SUCCEEDED	yava	Thu, 01 Dec 2016 18:44:12 GMT	Thu, 01 Dec 2016 18:44:48 GMT
0000000-161130113927475-oozie-oozi-W	demo	MR2	SUCCEEDED	yava	Thu, 01 Dec 2016 18:41:20 GMT	Thu, 01 Dec 2016 18:42:08 GMT
0000000-161129080054185-oozie-oozi-W	Hadoop	MR2	SUCCEEDED	yava	Wed, 30 Nov 2016 14:17:18 GMT	Wed, 30 Nov 2016 14:34:06 GMT
0000040-161116100421533-oozie-oozi-W	Ex11	MR2	SUCCEEDED	yava	Mon, 28 Nov 2016 20:25:25 GMT	Tue, 29 Nov 2016 14:19:09 GMT
0000039-161116100421533-oozie-oozi-W	Ex11	MR2	KILLED	yava	Mon, 28 Nov 2016 18:32:30 GMT	Tue, 29 Nov 2016 14:29:09 GMT
0000038-161116100421533-oozie-oozi-W	chiefs	MR2	SUCCEEDED	yava	Mon, 28 Nov 2016 18:27:20 GMT	Tue, 29 Nov 2016 14:18:24 GMT
0000037-161116100421533-oozie-oozi-W	Ex11	MR2	KILLED	yava	Mon, 28 Nov 2016 18:24:21 GMT	Tue, 29 Nov 2016 14:18:10 GMT
0000036-161116100421533-oozie-oozi-W	Ex11	MR2	KILLED	yava	Mon, 28 Nov 2016 18:24:04 GMT	Tue, 29 Nov 2016 14:18:09 GMT
0000035-161116100421533-oozie-oozi-W	Ex11	MR2	SUCCEEDED	yava	Mon, 28 Nov 2016 17:31:58 GMT	Mon, 28 Nov 2016 17:32:12 GMT
0000034-161116100421533-oozie-oozi-W	Ex11	MR2	SUCCEEDED	yava	Mon, 28 Nov 2016 17:20:56 GMT	Mon, 28 Nov 2016 17:21:10 GMT
0000033-161116100421533-oozie-oozi-W	Ex11	MR2	SUCCEEDED	yava	Mon, 28 Nov 2016 17:19:35 GMT	Mon, 28 Nov 2016 17:19:49 GMT
0000032-161116100421533-oozie-oozi-W	Ex11	MR2	SUCCEEDED	yava	Mon, 28 Nov 2016 17:17:57 GMT	Mon, 28 Nov 2016 17:18:11 GMT
0000031-161116100421533-oozie-oozi-W	Ex11	MR2	SUCCEEDED	yava	Mon, 28 Nov 2016 16:51:29 GMT	Mon, 28 Nov 2016 16:52:04 GMT
0000030-161116100421533-oozie-oozi-W	chiefs	MR2	SUCCEEDED	yava	Mon, 28 Nov 2016 16:32:06 GMT	Mon, 28 Nov 2016 16:32:41 GMT
0000029-161116100421533-oozie-oozi-W	Hadoop	MR2	SUCCEEDED	yava	Mon, 28 Nov 2016 15:29:27 GMT	Mon, 28 Nov 2016 15:30:07 GMT
0000028-161116100421533-oozie-oozi-W	Hadoop	MR2	SUCCEEDED	yava	Mon, 28 Nov 2016 15:11:38 GMT	Mon, 28 Nov 2016 15:12:17 GMT
0000027-161116100421533-oozie-oozi-W	Hadoop	MR2	SUCCEEDED	yava	Mon, 28 Nov 2016 14:48:27 GMT	Mon, 28 Nov 2016 14:49:34 GMT
0000026-161116100421533-oozie-oozi-W	Hadoop	MR2	SUCCEEDED	yava	Mon, 28 Nov 2016 14:47:42 GMT	Mon, 28 Nov 2016 14:48:52 GMT
0000025-161116100421533-oozie-oozi-W	Hadoop	MR2	SUCCEEDED	yava	Mon, 21 Nov 2016 15:24:42 GMT	Mon, 21 Nov 2016 15:25:57 GMT
0000024-161116100421533-oozie-oozi-W	tester	MR2	SUCCEEDED	yava	Mon, 21 Nov 2016 03:54:02 GMT	Mon, 21 Nov 2016 03:55:17 GMT
0000023-161116100421533-oozie-oozi-W	tester	MR2	SUCCEEDED	yava	Fri, 18 Nov 2016 12:24:32 GMT	Fri, 18 Nov 2016 12:25:42 GMT
0000022-161116100421533-oozie-oozi-W	tester	MR2	SUCCEEDED	yava	Fri, 18 Nov 2016 12:13:38 GMT	Fri, 18 Nov 2016 12:14:48 GMT
0000021-161116100421533-oozie-oozi-W	tester	MR2	SUCCEEDED	yava	Fri, 18 Nov 2016 12:07:27 GMT	Fri, 18 Nov 2016 12:08:37 GMT
0000020-161116100421533-oozie-oozi-W	encaes256	MR2	SUCCEEDED	yava	Fri, 18 Nov 2016 09:41:22 GMT	Fri, 18 Nov 2016 09:41:57 GMT

Global Big Data Conference



DISCOVER

INTEGRATE

MIGRATE

GOVERN

ANALYZE



The HDFS Browser and Data Viewer show the target file and its contents ..

You can also use the viewer window to manage all of your input and output data directly in HDFS..

The screenshot displays the IRI Workbench interface. The HDFS Browser window shows a file tree with the following structure:

- chiefs_some_left1
- chiefs_some_right
- claudia
- cm
- cmitra
- demo
 - chiefsMask
 - dept50_out
 - dept80_out
 - female_personal_in
 - male_personal_info
- donp_iri
- female_personal_info_
- java
- lib
- male_personal_info_er
- newyorkers.out
- NY_only.out
- oozie
- oozie-artefact
- oozie-oozi
- out_spark
- out_tez
- personal_info_decrypt
- personal_info_encrypt
- personal_info_name_d
- personal_info_name_e
- presOut
- pseudo_salOut

The Data Viewer window shows the contents of the file `/user/yava/demo/dept80_out/hgrid247-00000`:

Name	Size	Modified
._SUCCESS	1 KB	2016-12-01 14:58:54
hgrid247-00000	2 KB	2016-12-01 14:58:54

```
*****8594, Ellen, Abel, 47120, 0.*
*****8470, Sundar, Ande, 3225, 0.*
*****2514, Amit, Banda, 8202, 0.*
*****2093, Elizabeth, Bates, 2229, 0.*5
*****8917, David, Bernstein, 1117, 0.*5
*****1266, Harrison, Bloom, 19798, 0.*
*****5628, Nanette, Cambrault, 7379, 0.*
*****5712, Gerald, Cambrault, 47120, 0.*
*****7973, Louise, Doran, 7379, 0.*
*****4527, Alberto, Errazuriz, 92422, 0.*
*****6412, Tayler, Fox, 5394, 0.*
*****0747, Danielle, Greene, 1117, 0.*5
*****0916, Peter, Hall, 8398, 0.*5
*****0211, Alyssa, Hutton, 6953, 0.*5
*****7820, Charles, Johnson, 8202, 0.*
*****4077, Janette, King, 19798, 0.*5
*****3411, Sundita, Kumar, 9269, 0.*
*****5969, David, Lee, 3274, 0.*
*****1205, Jack, Livingston, 5524, 0.*
```

Global Big Data Conference



DISCOVER

INTEGRATE

MIGRATE

GOVERN

ANALYZE



Wizards for ...

New CDC Job

Data Targets
Specify the data targets and types of output. Your output fields need to be named the same as input fields to properly match; otherwise, use Target Field Layout.

If your output files contain a text description of the delta type, please select the field and enter that text in the text boxes. If Cumulative is selected, enter delta text separated by commas with no spaces in order of DELETE,EQUAL,INSERT,UPDATE.

Select output reports to produce

- Cumulative
Target: Cum.data
Format: DELIMITED
Metadata: metadata/Output.ddf
Delta: DELTA_FLAG, DELETE,EQUAL,INSERT,UPDATE
- Delete
Target: Delete.data
Format: DELIMITED
Metadata: metadata/Output.ddf
- Equal
Target: Equal.data
Format: DELIMITED
Metadata: metadata/Output.ddf
- Insert
Target: Insert.data
Format: DELIMITED
Metadata: metadata/Output.ddf
- Update
Target: Update.data
Format: DELIMITED
Metadata: metadata/Output.ddf

< Back Next > Finish Cancel

New SCD Job

Join Sources
To create a join condition, select a field to be matched from each Data Source, click a Join Type, and then click Create Condition (unless Unchecked Join).

Job Specification File
Define job specification file name, location, type of output, and SCD type.

Data Selection
Specify data sources, targets, format and metadata.

Data Mappings
Specify mappings for target. Place target fields in-line with matched source fields in table. Fill out combo boxes and tool fields as needed.

Pivot/Unpivot

New Pivot Job

Data Source
Select the input, key field, and pivot fields.

Source: C:/Eclipse/runtime-new/Pivot/Pivot.out
Format: DELIMITED
Metadata: metadata/pivot-year.ddf

Key Field: YEAR

Pivot Fields:

- YEAR
- DEPT100
- DEPT150
- DEPT200
- DEPT250
- DEPT300

Select All Select None

< Back Next > Finish Cancel

Slowly Changing Dimensions

Change Data Capture

Global Big Data Conference



DISCOVER

INTEGRATE

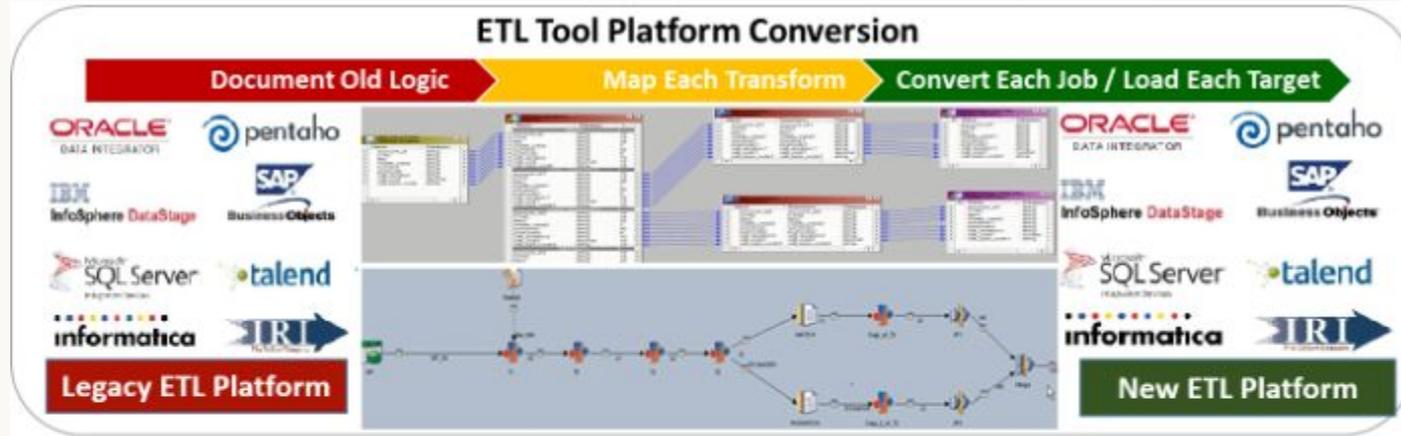
MIGRATE

GOVERN

ANALYZE



*With AnalytiX DS, ETL tool and SQL users can **convert** their existing data integration jobs to faster, simpler, and far less expensive Voracity workflows.*



Performance (like Ab Initio or Teradata)

Capability (like Informatica or DataStage)

DB affinity (like SSIS or ODI)

Eclipse ergonomics (like Talend)

Affordability (like Pentaho)



Global Big Data Conference



DISCOVER

INTEGRATE

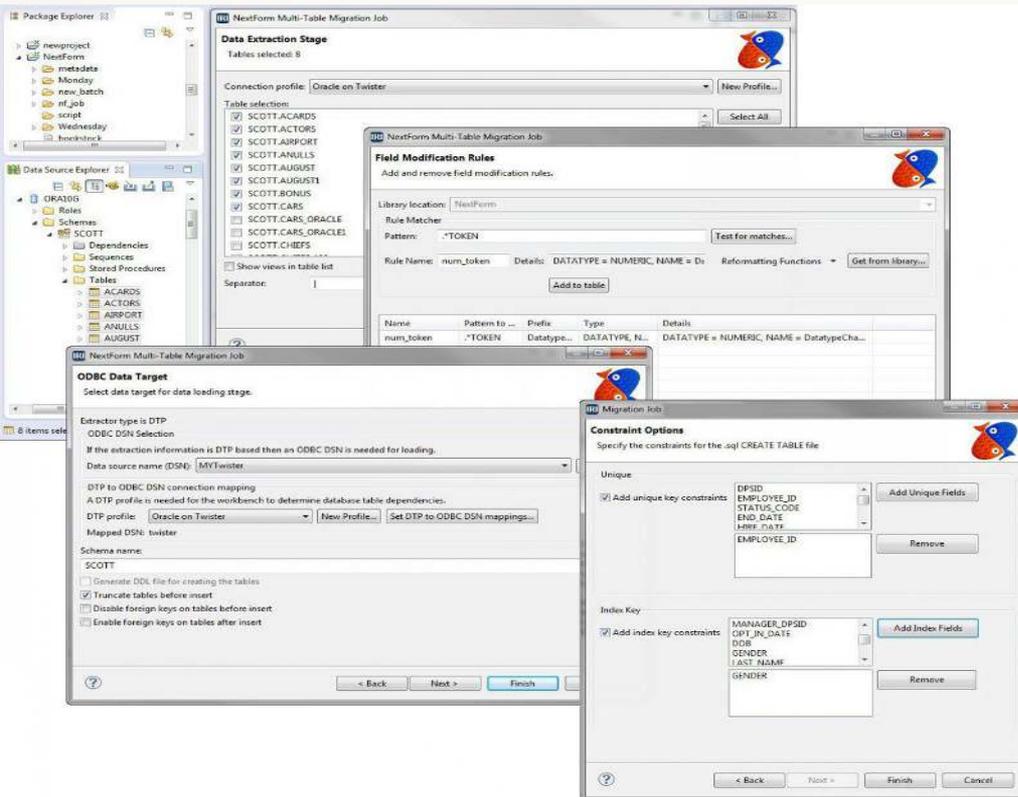
MIGRATE

GOVERN

ANALYZE



Voracity converts, replicates, and reformats data from mainframe datasets, relational and NoSQL databases, index and sequential files, dark data documents, and cloud apps.



- Change data types, record layouts, file formats, and endianness
- Migrate column values, layouts, and relationships (constraints) between DBs
- Copy or refresh data from one or more sources to one or more targets
- Federate, or virtualize, data by mashing up data from disparate sources and creating custom, ad hoc views

Global Big Data Conference



DISCOVER

INTEGRATE

MIGRATE

GOVERN

ANALYZE



- Connect and interact with **multiple sources** and targets, on-prem or cloud
- **Discover and classify** data in DB, flat-file, and dark-data (document) sources
- Mask **static or streaming** inputs, NoSQL DBs, and files in LUW, HDFS and S3
- Select from **12 masking categories** (e.g., encrypt, hash, pseudonymize, redact)
- **Address multiple** protections, targets and recipients all in one job, one I/O
- Apply consistent, cross-table masking rules for **referential integrity**
- Support **conditional security**, based on patterns, values, or ranges
- Specify target protections and formats in **Eclipse or portable** job scripts
- Integrate with **DB apps** via ODBC. Use .NET and Java SDK for dynamic masking
- Retain data **realism via FPE** and pseudonymization for testing, outsourcing
- **Mask during** big data ETL, migration, sub-setting, and BI/analytic jobs
- Log job and system runtime detail to XML audit files to **verify compliance**

Masking Features

Global Big Data Conference



DISCOVER

INTEGRATE

MIGRATE

GOVERN

ANALYZE



The screenshot displays the IRI Workbench interface. On the left, the Project Explorer shows a project structure with folders for 'Mongo', 'Project Dependencies', 'metadata', 'new_batch', and 'script'. The Data Source Explorer shows a 'Mongo Progress (MongoDB v. 3.0.6)' connection with a 'mydb' database containing a 'MYDB' schema with various tables like 'CUSTOMERS', 'CUSTOMERS_MASK', 'RESTAURANTS', etc.

The main window shows a configuration for 'MYDB_CUSTOMERS.fcl'. The configuration includes an INFILE section with the following rules:

```
/INFILE="MYDB.CUSTOMERS;  
/ALIAS=MYDB_CUSTOMER;  
/PROCESS=ODBC  
/FIELD=(PHONE, TYPE=V,  
/FIELD=(ID, TYPE=ASC,  
/FIELD=(STATE, TYPE=V,  
/FIELD=(NAME, TYPE=ASC
```

The OUTFILE section is configured as follows:

```
/OUTFILE="MYDB.CUSTOMERS;  
/PROCESS=ODBC  
/FIELD=(MASK_PHONE=PH  
/FIELD=(ID, TYPE=ASC,  
/FIELD=(ID, TYPE=ASC,  
/FIELD=(STATE, TYPE=V,  
/FIELD=(NAME, TYPE=ASCII, POSITION=5, SEPARATOR="|", EXT_FIELD="NAME")
```

Two data tables are shown for comparison:

PHONE [VARCHAR(4000)]	_ID [VARCHAR(24)]	ID [VARCHAR(4000)]	STATE [VARCHAR(4000)]	NAME [VARCHAR(4000)]
8514532145	5630BC2D259D18...	12409	OHIO	REID
9654125893	5630BC2D259D18...	85460	MAINE	CHARLES
9641258637	5630BC2D259D18...	95364	GEORGIA	FOSTER
2156354789	5630BC2D259D18...	15634	CALIFORNIA	TIM
3216874517	5630BC2D259D18...	85475	IDAHO	FRANK
5126987456	5630BC2D259D18...	45214	TEXAS	BOB
4582147965	5630BC2D259D18...	32567	NEVADA	DAVID
2145896732	5630BC2D259D18...	14563	NEW YORK	ISABELLE
5412395475	5630BC2D259D18...	25469	WASHINGTON	PAUL
5214596324	5630BC2E259D18...	45625	COLORADO	RON
<new row>				

PHONE [VARCHAR(4000)]	_ID [VARCHAR(24)]	ID [VARCHAR(4000)]	STATE [VARCHAR(4000)]	NAME [VARCHAR(4000)]
851*****	5630BC2D259D18...	12409	OHIO	REID
965*****	5630BC2D259D18...	85460	MAINE	CHARLES
964*****	5630BC2D259D18...	95364	GEORGIA	FOSTER
215*****	5630BC2D259D18...	15634	CALIFORNIA	TIM
321*****	5630BC2D259D18...	85475	IDAHO	FRANK
512*****	5630BC2D259D18...	45214	TEXAS	BOB
458*****	5630BC2D259D18...	32567	NEVADA	DAVID
214*****	5630BC2D259D18...	14563	NEW YORK	ISABELLE
541*****	5630BC2D259D18...	25469	WASHINGTON	PAUL
521*****	5630BC2E259D18...	45625	COLORADO	RON

MongoDB Masking

Global Big Data Conference



DISCOVER

INTEGRATE

MIGRATE

GOVERN

ANALYZE



Define once, deploy everywhere

Name	Size	Modified
female_personal_info_encrypted		11/17/2016 07:59:00
male_personal_info_encrypted		11/17/2016 07:59:29
chiefs.txt	2 KB	11/16/2016 15:50:55
personInformation	1 KB	11/16/2016 15:48:21
personInformation2	1 KB	11/21/2016 10:20:54
purchases	1 KB	11/16/2016 15:23:50

Name	Size	Modified
female_personal_info_encrypted		11/17/2016 07:59:00
male_personal_info_encrypted		11/17/2016 07:59:29
chiefs.txt	2 KB	11/16/2016 15:50:55
personInformation	1 KB	11/16/2016 15:48:21
personInformation2	1 KB	11/21/2016 10:20:54
purchases	1 KB	11/16/2016 15:23:50

Masking in Hadoop

Global Big Data Conference



DISCOVER

INTEGRATE

MIGRATE

GOVERN

ANALYZE



- Create synthetic but realistic **random and random-real** test data simultaneously
- Improve **DB prototypes**, application quality, benchmarking, and devops
- Leverage DDL, production file, and/or custom metadata
- Preserve structural and **referential integrity**
- Produce data in any type, structure, volume, value range, and “if” condition
- Synthesize **composite values** and custom (master) data formats
- Generate computationally valid and invalid NID, SSN, or CC#
- Set and graph test data **value distributions** (linear, normal, random, etc.)
- Apply common attribute rules (e.g., lookups) for pattern-matched field names
- **Filter, transform, and pre-sort** test data as you generate it
- Write loader metadata, and perform the loading, automatically
- Build test flat-file and custom detail and summary reports
- **Subset and mask** databases automatically as an alternative approach
- Use Java SDK functions to generate test data in apps and Hadoop

TDM Features

Global Big Data Conference



DISCOVER

INTEGRATE

MIGRATE

GOVERN

ANALYZE



From its one Eclipse IDE (IRI Workbench) Voracity supports multiple analytic approaches ...

Voracity Analytic Option 1: Embedded BI

Unlimited 2D reporting
in custom-formatted,
detail and summary
files, XML, HTML, etc.

The screenshot displays the IRI Workbench interface with several panes:

- Project Explorer:** Shows a project structure with folders like 'InputFiles', 'Jobs', 'Targets', and 'TradingAhtml'.
- nyse-a:** A table of stock data with columns for company name, AOS, AGE, and others.
- buys.csv:** A table with columns for Shares, Symbol, and Client, listing names like Bill Gates and Warren Buffet.
- stockjoin.scl:** A script defining data sources and joins.
- Console:** Shows the execution of a SortCL job, displaying a table of client trades with columns for Client, Symbol, Shares, LastTrade, and Shares*LT.
- HTML produced by:** A summary report listing client names and their associated stock symbols (e.g., Stephen Covey - HBC, Richard Branson - HIG).
- TradingAhtml:** An XML report showing detailed trade information for clients like Jack Welch and Michael Bloomberg.



Global Big Data Conference



DISCOVER

INTEGRATE

MIGRATE

GOVERN

ANALYZE



Voracity Analytic Option 2: BIRT Integration

Prepare and present
data simultaneously
from an "IRI Data
Source" in BIRT

The screenshot displays the IBM Business Intelligence Reporting Tools (BIRT) interface. The main window shows a report titled "Bank of America Moving Averages" with a line chart. The chart has a y-axis ranging from 70 to 95 and an x-axis with 8 data points. The data points are: (1, 85), (2, 80), (3, 90), (4, 75), (5, 80), (6, 85), (7, 80), (8, 85). The chart is titled "Bank of America Moving Averages".

The interface includes a Project Explorer on the left, a Data Source Explorer at the bottom left, and a Console at the bottom right. The Console shows the following data:

Date	Value
01/02/2013,12.05,12.15,11.9,12.03,236021300	
01/03/2013,12.01,12.05,11.80,11.96,157149600	
01/04/2013,11.97,12.11,11.93,12.11,132605600	
01/07/2013,12.15,12.2,12.12,12.09,201403500	
01/08/2013,12.09,12.1,11.89,11.98,168464600	
01/09/2013,11.87,12.11,11.33,11.43,336167600	
01/10/2013,11.61,11.81,11.54,11.78,199964900	



Global Big Data Conference



DISCOVER

INTEGRATE

MIGRATE

GOVERN

ANALYZE



Voracity Analytic Option 3: Cloud Dashboard

Leverage drill-down,
browser-based
dashboard applications
like this one from
NextCoder

The screenshot displays the IRI Development platform interface. The top window shows a workflow diagram for 'Sort_CDR4.scl' with steps like 'Sort_Prefix.scl', 'Sort_Tariff.scl', 'Sort_CDR2.scl', and 'Sort_CDR4.scl'. The middle window shows SQL Results for a query: 'SELECT * FROM "SCOTT"."XL4"'. The bottom window shows a dashboard titled 'Calls By Trunk Out' with a horizontal bar chart and call time information.

START_TIME	GLOBAL_CALL_ID	DURATION	CALLING_NUMBER	CALLED_N
201503200...	1001	20	0818000000	0817000000
201503200...	1002	20	0817000000	0818000000
201503200...	1002	20	0817000000	0818000000

Calls By Trunk Out

Legend: Grouped, Stacked, Amount (K), Duration

Trunks: TXL1, TXL2, TTLK1, TSEL1

Earliest Call: 05:00

Last Call: 22:00



Global Big Data Conference



DISCOVER

INTEGRATE

MIGRATE

GOVERN

ANALYZE



Voracity Analytic Option 4: Splunk Add-On

Prepare data you need
to index ad hoc, with a
Voracity job launched
from Splunk

The screenshot shows the Splunk web interface. The top navigation bar includes 'splunk>', 'Apps', 'Administrator', 'Messages', 'Settings', 'Activity', and 'Help'. The main content area is titled 'Add Data' and has a progress bar with three steps: 'Select Forwarders', 'Select Source', and 'Done'. Below this, there are several search results tabs: 'System, Processor, Service information about this machine', 'Local Windows network', 'Local Windows print', 'Scripts', 'hello', 'iri', 'IRI_Voracity', and 'Powershell v3 Modules'. The 'iri' tab is selected, showing a search for 'host=lava' with 1 event. The search results are displayed in a table with columns for 'Time' and 'Event'. The event details are as follows:

i	Time	Event
>	3/14/16 3:40:18.000 PM	"Sara", "Tiemann", "Gadsden", "AL", "*****1643", "694-06-0760" "James", "Wadsworth", "Prattville", "AL", "*****7526", "498-97-75" "Bonnie", "Simmons", "Arkadelphia", "AR", "*****6221", "189-82-17" "Amanda", "Bess", "Fort Smith", "AR", "*****1643", "281-55-5360" "Dolores", "Miles", "De Queen", "AR", "*****2418", "061-90-2361"

Selected Fields: host 1, source 1, sourcetype 1. Interesting Fields: host = lava, source = iri//IRI TEST, sourcetype = iri.



Global Big Data Conference



DISCOVER

INTEGRATE

MIGRATE

GOVERN

ANALYZE



Voracity Analytic Option 5: Data Blending

Prepare CSV, XML or table subsets to reduce time-to-display 2-20X, along with data quality, privacy, and storage



Global Big Data Conference



DISCOVER

INTEGRATE

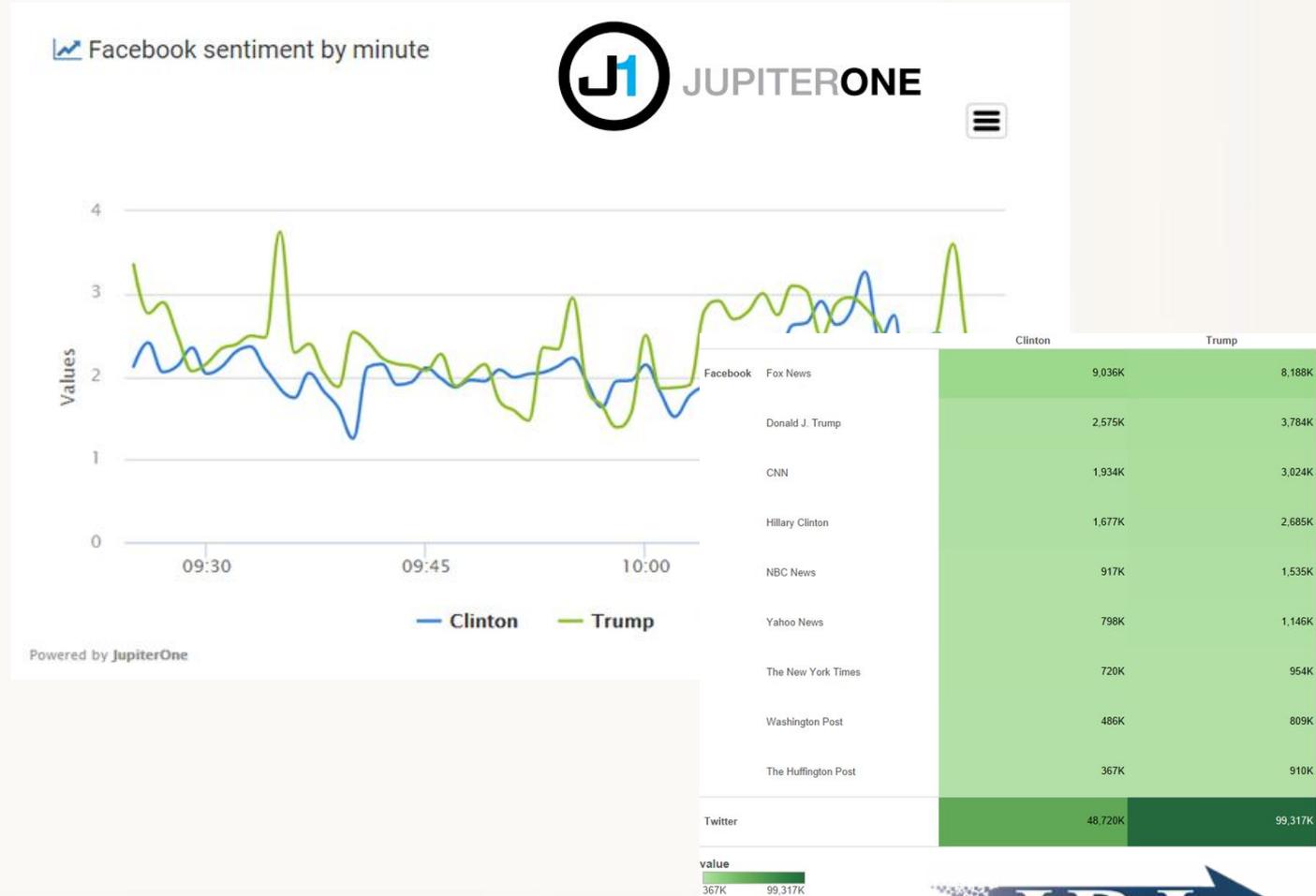
MIGRATE

GOVERN

ANALYZE



Voracity Analytic Option 6: Big SM Streams



Leverage advanced text and social media analytic engines with NLP and Kafka support



Global Big Data Conference



DISCOVER

INTEGRATE

MIGRATE

GOVERN

ANALYZE



Data Curation

Profile & Acquire

Discover and extract data and metadata in disparate sources. Define custom structures, mask formats, and build test data.

Cleanse & Unify

Filter, enrich, scrub and standardize data in multiple sources. Select, fuzzy-search, and merge reference data into master tables and values.

Process & Provide

Integrate, migrate, govern, and analyze data in the same job and I/O pass. Visualize and feed test or real targets in any format.

Protect & Audit

De-ID data at the field level as you acquire, transform, report, or franchise. Encrypt, hash, pseudonymize, redact, tokenize, etc.

Express & Predict

Aggregate, cross-calc, and format data in detail, summary and trend reports, or, hand-off results to your analytic tool or BIRT charts in memory.

Convert & Replicate

Migrate legacy databases, or files and data types -- or specify new target record layouts -- in copies, or subsets, of data in any format or schema.

Publish & Share

Federate, save, or populate multiple targets at once, and connect to them and their metadata in secure repositories for change tracking, etc.

Global Big Data Conference



DISCOVER

INTEGRATE

MIGRATE

GOVERN

ANALYZE



Voracity Uses

Retail

Micro-target customers

Use Voracity to segment purchase groups for targeted marketing, and to create holistic, unified views of each customer that help you customize service and build loyalty.

Leverage Consumer Psychology

Use Voracity to integrate consumer behavior and sentiment data against seasonal, regional, weather, and other factors, and mine it with regression analyses that reveal trends.

Price Smarter

Use Voracity to integrate preference and pricing data from retail data brokers, public data, your own pricing history, and competitive research.

Global Big Data Conference



DISCOVER

INTEGRATE

MIGRATE

GOVERN

ANALYZE



Voracity Uses

BFSI

Assess Credit Risk

Use CoSort and Hadoop engines in Voracity to blend traditional credit data with sources like utility bill and rental payments to improve score accuracy, facilitate lending, marketing, etc.

Optimize Loan Performance

Use Voracity to blend and prepare internal and external data points (borrower history, industry repayment stats, social/market forces, etc.) for visual analytics on risk factors vs. loan rates.

Expose Insurance Fraud

Use Voracity to rapidly sort, filter, and expose claim data outside normal parameters to identify suspicious behavior, and feed it to visualization and notification apps in the same IDE.

Global Big Data Conference



DISCOVER

INTEGRATE

MIGRATE

GOVERN

ANALYZE



Healthcare

Voracity Uses

Improve Treatment Outcomes

Flow IoT data through slowly changing dimension or change data capture processes in Voracity to compare patient data with diagnostic values to spot, alert, and correct for abnormalities.

Individualize Drug Therapies

Rapidly integrate genetic data into single-node-type networks, gene-set libraries, and bi-partite graphs to help reveal new relationships between patient genes, drugs and phenotypes.

See the Whole Patient

Use Voracity' search, join, consolidate, and masking features to unify and de-identify patient information from family, provider, demographic, diagnostic and treatment data silos.

Global Big Data Conference



DISCOVER

INTEGRATE

MIGRATE

GOVERN

ANALYZE



Energy & Transport

Conserve & Troubleshoot

Use the IoT edge aggregation and hub analytics in Voracity on smart meter and thermostat data to identify peak uses, or on grid sensor and weather data to re-route power, inspect, repair, etc.

Improve Traffic Flow

Combine data from street cameras and sensors, cell phone apps and weather data in Voracity and feed it directly into BIRT or BIRT-connected Integeo geospatial reports to warn drivers.

Optimize Fleet Performance

Use IoT analytic and alerting features in Voracity to predict and prevent equipment failures, and its DW/BI prowess against historic O&D and pricing data to maximize passenger revenues.

Voracity Uses

Global Big Data Conference



DISCOVER

INTEGRATE

MIGRATE

GOVERN

ANALYZE



Voracity Uses

Telco & Media

Monetize Calls & Clicks

Use Voracity to process CDRs and clickstream data for billing and analytics, and to sell that data to marketing affiliates and others who can permissibly use it.

Anticipate Spending Trends

Use Voracity to extract string and pattern-matching values from social data from Hubspot, etc., and munge it with transaction and demographic data to identify and predict content preferences.

Throttling & Enforcement

Use Voracity to identify excessive bandwidth usage or illegal behavior from network traffic and web logs, and tie it to analytic and notification mechanisms in the same IDE.

Global Big Data Conference



DISCOVER

INTEGRATE

MIGRATE

GOVERN

ANALYZE



Voracity Uses

Reliance Communications (RC) is broadband and telco company in india with 110M subscribers. To meet daily SLAs in billing and analytics for wireless (mobile) and global (landline) segments, RC must process and report on hundreds of millions of call detail records (CDRs) every day.

RC uses 64-bit Solaris servers and Oracle. The CDRs come from binary switch data mediated into flat files that the CoSort engine in Voracity transforms *before* DataStage ETL & BOBJ reports.

“Prior pilots failed from slow and inaccurate results, and SLAs were missed as call volume grew. After Voracity jobs transformed flat files in the 60GB range, the processing bottleneck disappeared, and our analytic results were always accurate.”



Global Big Data Conference



DISCOVER

INTEGRATE

MIGRATE

GOVERN

ANALYZE



Voracity Uses

DataBase Technologies (DBT) in Parsippany, NJ builds and maintains VLDB CRMs for ADP, Verizon, Merrill Lynch, Seagrams, and Universal Studios.

DBT integrates 350M transaction records per day, joining them to files up to 100M rows each, and accumulating the data over time for analysis. Their first 350GB dataset took over two days to load, so it had to be pre-sorted.

"It's fun to watch the system performance monitor and see all those processors working in the high 90 percentages and the disks utilizing the fast data rates you pay for."

Voracity filter, sort, and join operations, were 10x faster than those in MS SQL Server 9.5 minutes versus 98 @350GB.



Global Big Data Conference



Learn and Share

IRI.com [IRI blog](#)



[IRI Voracity Data Management Group on LinkedIn](#)

