CHALLENGES OUR CLIENTS ARE FACING

IN THE PAST 15 YEARS, EPAM HAS WORKED WITHIN A VARIETY OF INDUSTRIES DRIVEN BY DATA SUCH AS FINANCE, MEDIA AND HEALTH CARE. WITH THE ADVENT OF NEW DATA TECHNOLOGIES, EPAM ALSO HAS SEEN HOW COMPANIES ARE BECOMING MORE AGILE IN THE FACE OF DATA CHALLENGES.

With more systems freely allowing data to be extracted and shared, the healthcare industry is faced with the following challenges:

- Health data is fragmented and often unstructured
- Non-uniformity of clinical data formats, code sets and identification
- Collection and analytics of consumer behavior data and incentive information increases complexity of data challenges
- Exponentially growing data volume
- Scaling rules-driven implementations efficiently for large populations to support transactional and analytics use cases
- Software and Hardware costs are capital intensive

To address these issues, most organizations are considering new technologies and ways of thinking. EPAM has created a Healthcare Big Data Accelerator to help its partners quickly implement Big Data operations using cutting edge data analytics platform and components.
## HEALTHCARE TRENDS WE SEE INFLUENCING INNOVATION

<table>
<thead>
<tr>
<th>HEALTHCARE TRENDS</th>
<th>IMPLICATIONS ON ANALYTICS</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>1 CONSUMERISM</strong></td>
<td>Better engagement and outcomes require key capabilities to analyze consumer behavior, personalize experiences, measure and manage incentives</td>
</tr>
<tr>
<td><strong>2 VALUE BASED CARE</strong></td>
<td>- Increased use of clinical data to measure quality of care accurately in timely fashion</td>
</tr>
<tr>
<td></td>
<td>- Manage transitions of care and aggregate data beyond organizational boundaries</td>
</tr>
<tr>
<td></td>
<td>- Population health management support for provider organizations</td>
</tr>
<tr>
<td><strong>3 INCREASED VELOCITY AND VARIETY OF HEALTHCARE DATA</strong></td>
<td>- The type and number of data sources (variety) plus frequency of analysis (velocity) is increasingly driven by two trends mentioned above</td>
</tr>
<tr>
<td></td>
<td>- The new trends necessitate complex and agile analytics delivery models that can adapt</td>
</tr>
</tbody>
</table>
HOW EPAM CLIENTS ARE REALIZING VALUE FROM OUR SOLUTION

1. **DECREASE TIME TO INSIGHT**
   - Move away from schema first, big scale engineering effort to a more business use-case driven approach
   - Rigid data schemas are difficult to change without impact on downstream systems

2. **REDUCE COST**
   - Software, storage and processing
   - Elasticity

3. **EXPLORE IMMEDIATE AVAILABILITY OF DATA FOR DISCOVERY AND EXPLORATION**

4. **SHARED TECHNOLOGY IMPLEMENTATION AND RESOURCES BETWEEN SYSTEMS OF ENGAGEMENTS AND ANALYTICS APPLICATIONS**
   - Speed layer manages event-based and streaming data with incremental updates
   - Batch layer manages population level processing using the shared code base

5. **NOW-CASTING USING NEAR REAL-TIME DATA AT POPULATION OR INDIVIDUAL LEVEL**
   - Parallel computing and rules execution allows now-casting for millions of patient records
For each required scenario, shared service components are used in purchased 3rd party data sets.

**Reference**
- Census
- Purchased 3rd Party Data Sets
- CMS Data

**PATIENT**
- Clinical & Demographic Data HDFS

**Clinical**
- Care Gaps
- Personalized Treatment Plans
- Risk and Predictive Models

**Device**
- Trackers
- Portal Data

**Refined Area**
- Diagnosis
- Medications
- Allergies
- Immunizations
- Problems
- Lab Data
- Admission and Discharge

**Population level statistics, HEDIS scoring**

**REPORTING AND ANALYTICS**
- Exploration
- BI Analytics

**SYSTEMS OF ENGAGEMENT**
- Clinical Apps
- Application Services
- RDBMS
HL7 ADAPTERS IN BIG DATA

DATA STRUCTURES

Challenge: Data sources provide flat files in HL7 format that are not immediately useful for data processing.

Solution: Create Avro models that can be queried by interested parties.

ACCESS TO DATA

Challenge: Data about a given population is split across multiple flat files that contain hundreds of thousands of rows. However, not all data is accessible for processing due to the complexity it introduces to existing ETL processes.

Solution: Generate required number of views on the HL7 model to allow for querying of interested data elements. For CCD records, allow users to directly query the data inside CCD file using a combination of SQL and XPath queries.

HL7 OVERVIEW

- HL7 Transactions - Uses Open Source HAPI API
- CCD – Transform XML file in optimized Parquet files

Prepare Avro Models from HL7 Input

Query Clinical Data

- Query HL7 transactions through Avro files using SQL Like language
- Query CCD files using a combination of SQL and XPath
# TABLES AND QUERIES VIEWS

```sql
SELECT * FROM LAB_DGN LIMIT 100
```

<table>
<thead>
<tr>
<th>Table Name</th>
<th>Browse Data</th>
<th>Browse Data</th>
<th>Browse Data</th>
<th>Browse Data</th>
<th>Browse Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>lab_dgn</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>lab_messages</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>lab_msg</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>lab_ordr</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>lab_rst</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>lab_dgn.lab_msg_id</th>
<th>lab_dgn.snp_num</th>
<th>lab_dgn.dgnns_cd</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>1</td>
<td>W</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>Jk</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>5i</td>
</tr>
<tr>
<td>1</td>
<td>4</td>
<td>K</td>
</tr>
<tr>
<td>2</td>
<td>5</td>
<td>k</td>
</tr>
<tr>
<td>3</td>
<td>6</td>
<td>F</td>
</tr>
<tr>
<td>4</td>
<td>7</td>
<td>12</td>
</tr>
<tr>
<td>5</td>
<td>8</td>
<td>21</td>
</tr>
<tr>
<td>6</td>
<td>9</td>
<td>21</td>
</tr>
<tr>
<td>7</td>
<td>10</td>
<td>41</td>
</tr>
<tr>
<td>8</td>
<td>11</td>
<td>q</td>
</tr>
<tr>
<td>9</td>
<td>12</td>
<td>v</td>
</tr>
<tr>
<td>10</td>
<td>13</td>
<td>y</td>
</tr>
<tr>
<td>11</td>
<td>14</td>
<td>z</td>
</tr>
<tr>
<td>12</td>
<td>15</td>
<td>A</td>
</tr>
<tr>
<td>13</td>
<td>16</td>
<td>B</td>
</tr>
<tr>
<td>14</td>
<td>17</td>
<td>C</td>
</tr>
</tbody>
</table>
EMPI INTEGRATION INTO HADOOP ARCHITECTURE

**DATA LOCALITY**

**Challenge:** Hadoop favors data localization to scale efficiently while EMPI solutions are inherently central to guarantee unique patient identification.

**Solution:** Use a separate identification stage to enhance the metadata of records and store it in HBase before the parallel clinical processing pipeline is executed.

**SCALABILITY**

**Challenge:** Patient record must be identified individually across all data sets, requiring an inherently parallel architecture for EMPI resolution.

**Solution:** Minimize the number of ID resolution calls in a parallel ecosystem by utilizing the distributed cache solution in Hadoop to store patient lookup information.

**DATA QUALITY**

**Challenge:** Keeping the EMPI data quality high as data input volume and variety increases over time.

**Insights:** A separate maintenance and update cycle is needed to refresh the EMPI data. To facilitate this in the Hadoop environment, by utilizing a native Hadoop DB.

*EPAM Accelerator is pre-configured to work with OpenEMPI. It can also be configured to integrate with existing enterprise MPI solutions.*
Clinical Data Files are stored in the Raw area.

2. Mappers group data by source patient identifier such as SSN, ID, etc.

3. All relevant records of a patient are grouped together.

4. EMPI System is called to resolve source patient identifier (such as SSN) to target Patient Identifier – only calling MPI system once per patient per source.

5. Results are stored in HBase for clinical processing.
CLINICAL DATA TRANSLATION

INSIGHTS

COMMON DATA REPRESENTATION

Challenge: Writing general rules against non-homogenous datasets is not feasible.

Solution: Create a unified, normalized view that contains standard coding scheme that rules can refer to.

NON-UNIFORM CODING

Challenge: Clinical data contains codes using different coding schemes like CPT, ICD9, NDC, etc.

Solution: Use a common rich standards based terminology and concept dictionary service like HDD (3M).

TRANSLATION SERVICE COVERAGE

Challenge: Clinical data may contain non-standard coding schemes that may not be covered by existing HDD translation services.

Solution: Provide extensible code lookup mechanism by way of abstraction for introducing new dictionaries.

CODING TRANSLATION PERFORMANCE / SCALABILITY

Challenge: HDD Access provides service that use complex logic to translate codes. Non-local service calls conflicts with Hadoop architecture.

Solution: Deploy HDD service and database to each data node to localize code lookups for data translation.

BIG DATA ACCELERATOR CODING TRANSLATION

In order to localize HDD access calls, HDD service and database is deployed to all data nodes.
INSIGHTS

BUSINESS FRIENDLY RULES LANGUAGE

**Challenge:** Business requires a possibility to define business rules without deep technical background.

**Solution:** Drools as a business rules engine that supports business friendly rules engine, Excel spreadsheet for decision map, extendable DSL language.

PARALLEL COMPUTATION

**Challenge:** Thousands of rules are applied to millions of patient profiles - causing millions of calculations. And while healthcare companies should continuously analyze patient information to provide proper treatment, sometimes even daily updates are not enough to provide the result the doctor needs immediately during a patient visit.

**Solution:** Hadoop provides an extremely efficient functionality to perform computation on massive amounts of data in parallel. Spark Streaming provides the possibility to perform calculations in near real-time fashion. Drools is a Java based framework that allows native integration with Big Data technologies (MapReduce, Spark, Spark Streaming).

DROOLS WORKBENCH

**Challenge:** Business users require an easy tool to create, validate and deploy business rules.

**Solution:** Drools Workbench provides powerful web UI for rule management. Native integration with Maven allows integration of business rules into development life cycle including AB testing.
SAMPLE RULE (DROOLS WORKBENCH)
# HOW WE ENSURE DATA SECURITY

## KERBERIZED ENVIRONMENT
- HDP or Cloudera distributions allow enable Kerberos
- Integration with Active Directory

## ENCRYPTION
- Encryption at Rest
  - HDFS Level
  - OS Level
  - Hardware Encryption
- Encryption in Motion
  - Hadoop based
  - Network Based
  - Secured Perimeter

## AUDITING
- Cloudera Navigator
- Apache Ranger

## CODING ROLE BASED ACCESS
- Combination of Two Dimensional access
- Per Data Source access
- Security classification

---

**TO LEARN MORE ABOUT SCHEDULING A COMPLIMENTARY WORKSHOP**

CONTACT US AT:  
JOHN_JUDGE@EPAM.COM

For more information, PLEASE VISIT EPAM.COM

---

24 West 25th Street, 5th Floor  
New York, NY 10010  
P: +1-267-759-9000  |  F: +1 267 759 8989

222 Kearny Street, Suite 308  
San Francisco, CA 94108

© 1993-2015 EPAM. All Rights Reserved.