MIT SOCIOTECHNICAL SYSTEMS RESEARCH CENTER

Hadoop-Based Data Exploration for the Healthcare Safety-Net – Technical & Sociocultural Challenges to Big Data Usability

David Hartzband, D.Sc.
Research Affiliate, SSRC, MIT
&
Director, Technology Research
RCHN Community Health Foundation

Path2Analytics Project

- Path2Analytics is a project developed & supported by the RCHN Community Health Foundation (rchnfoundation.org) that deploys a Hadoop-based analytic stack into Federally Qualified Health Centers & works with the Exec & IT Staffs of the organizations so that the center can use this type of analysis as a strategic decision aid after the end of the sponsored effort.

MIT SOCIOTECHNICAL SYSTEMS RESEARCH CENTER

Data as an Asset

- Two dimensions:
  - Socio-organizational
    - Use of data is not an IT function! It must be an integral & integrated aspect of health center strategy.
    - Analysis of data must be done in a strategic context
    - Thinking about data must permeate all discussion so that all decisions are, in part, data-driven decisions
    - Data is not just health center data, but all available data
  - Technical
    - Management of data as well as acquisition & adoption of analytic tools must be a long-term function, regardless of where it is located (IT, Strategy, Mgmt. Staff etc.)
    - Analytic function requires personnel, training, dedicated hardware & software
    - Tools & type of analysis must align with health center capabilities & strategy, not external opinion or fad

MIT SOCIOTECHNICAL SYSTEMS RESEARCH CENTER

Project Status

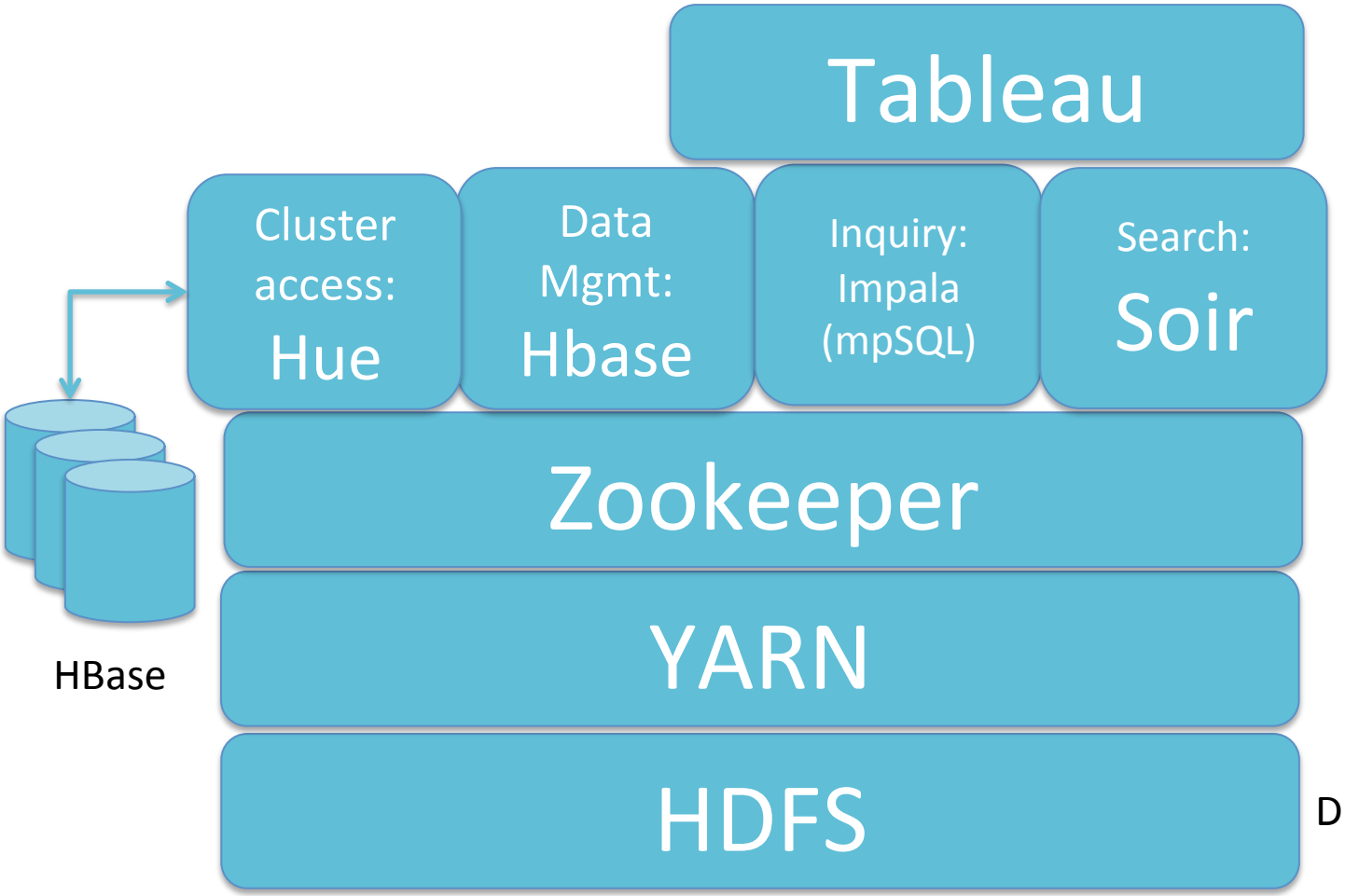
- Currently three health centers active:
  - Analytic stack deployed at two centers with preliminary results
  - Work done with Staffs at both health centers for definition of inquiry & use of analytic capability
  - Planning work done at one center
  - Starting discussions with a PCA with 24 centers in their HCCN
- Overall: 3 centers, 45 sites, ~100K patients/year
- Two of the health centers are urban
  - Both with 13-18 sites,
    - one with ~25K patients/yr, ~ 200K encounters/yr
    - one with ~55K patients/yr, ~450K encounters/yr
- One of the health centers is rural
  - 14 sites, ~15K patients a year, <50K encounters/yr
    - Some sites have <10 encounters/week

MIT SOCIOTECHNICAL SYSTEMS RESEARCH CENTER

Proposed Technical Approach

- Preliminary characterization of data using existing reporting/BI system
  - Can be used to determine if normalization/ETL is needed
- Use of existing data extracts/warehouse if available
  - Analysis of data quality needs to be done
- Layering of existing data into open source analytic stack (Hadoop-based, Cloudera open-source distro)
- Use of Yarn/MapReduce 2 to generate results (R), &/or
- Use of Hbase for data management
  - All data stays within health center's security perimeter
  - Only results leave perimeter
- Use of Hive/Impala to do query
- Use of Tableau to visualize raw data

Deployed Analytic Stack



Deployment Configuration

App User Provider, Front/Back Office

PCA Deployment:
- 4 node Linux cluster
- 4 cores each
- 3.5 GHz processors
- 64 GB memory
- 8 TB storage

Analytics Server
- Single node
- Unbuntu Linux (14.1)
- virtual, equiv 4-core, 3-3.5 GHz
- 16GB memory
- 1TB Storage

Analytics User

Financial data managed & analyzed in a separate system,… will be integrated into analytic stack…

Cloudera Express Hadoop Distro

Analytics Server

App Server EHR, PM…

Report Server

Data Warehouse

Data Volumes

- Urban CHC 1 – nightly extract to Report Server, ~250GB/yr, three usable data years (2012-2014)
- Urban CHC 2 – nightly extract to Report Server HCCN-based), separate Data Warehouse (6 data years, ~approx 1 month lag), ~2.2TB
- Urban CHC3 – nightly extract to remote repository (HCCN-based), ~400GB/yr
- Rural CHC 1 – read-only access to EHR DB, ~65GB/yr

Data estimates do not include images or financial

Summary of Anomalous Results - 1

- Omission, commision & migration errors
  - Multiple different descriptions for ICD-9 codes
    - In one instance, a single code had 29 entirely different descriptions associated with it
    - Thought to be the result of a migration done for the 2009-2011 data years moving from Mysis to NextGen
  - Many demographic & clinical fields left blank
  - Clearly incorrect data entered for several clinical fields
    - BMI ranges from 5-301,000
    - A women 5'1", 325 lbs has a BMI of 61.4, no one alive has a BMI of over 100, let alone 300,000
    - Similar incorrect ranges for blood pressure & static glucose

Summary of Anomalous Results – 2

- Non-Alignment of Clinical & Financial data
  - In two of three cases, clinical & financial data is not only entered, managed & analyzed in different software systems, there is no technical link or only a tenuous link (key, common data) between them
  - Financial data not organized by patient or encounter but rather summarized by location, time period, payer type… link to patient &/or encounter can be analyzed & inferred, but not direct
  - In all cases, staff responsible for each data type is not aware of the characteristics of the other system or data, & does not appear interested
  - In one case, financial data appears to have an encounter key, but this does not seem to be tied to a clinical encounter, we are still exploring this data

Summary of Anomalous Results - 3

- Patient Count Discrepancies
  - Patient counts were initially calculated as number of unique patient IDs that had an encounter with an associated diagnosis per year
  - In one case = 25,353, CHC staff including CMO stated this was "way too high"
    - CHC calculated patients as number of unique patient IDs with "billable medical encounters" with associated diagnosis/yr minus dental patients that did not get primary care at CHC = 19011
    - Billable medical encounters does not count:
      - Most behavioral health, optometry, medical counseling (nutrition, diabetes, etc.) screening procedures, nursing visits…
    - CHC agreed to add all patients except unique dental back into count = 21054
  - Diagnosis percentages & comorbidities were calculated using both numbers (percentages not significantly different)

Summary of Anomalous Results - 4

- Cultural & Training Bias Errors
  - Health Center populations are not healthier than the U.S. population in general
    - CDC FastStat data used to compare percentage of diagnoses from 2 cases
    - Many diagnosis types substantially underdiagnosed vs. U.S. data
    - This is especially true of obesity & heart disease with hypertension, diabetes & behavioral health also underdiagnosed

	CHC %	U.S. %	CHC/U.S%
hypertension	~20%	~30%	67%
diabetes	~6%	~9%	67%
obesity	~4%	~35%	11%
heart disease	~3%	~11%	25%
behavioral	~17%	~25%	68%

Implications of Anomalies - 1

- Omission, Commision & Migration
  - In one case, data from years 2009-2011 was not usable
  - In second case (no migration) there were many fewer errors & only one data year (2009) was not usable
  - In the case of incorrect data entered, analysis is not reflective of actual population & extreme values skew results
- Financial & Clinical Nonalignment
  - We are still exploring whether financial & clinical can be related at an encounter &/or patient level at two of the health centers
  - If not, then any analysis of cost per parameter (diagnosis, patient, comorbidity cluster, location, provider etc.) will not be possible
  - Third health center appears to have encounter based linkage, will need to determine how typical either case is in general at health centers

Implications of Anomalies - 2

- Patient Count
  - Would like to have a simple (non-exclusionary) algorithm for calculating patient counts
  - Current algorithm for one urban center is complex, exclusionary & cannot be used at other centers (because they do have such a high percentage of dental patients or do not provide dental care, they do not agree on using only billable encounters, they also count some "non-medical" encounters,…)
  - If every center has a unique way of defining "patients", or any other core term (encounter, provider, outcome, service, cost), no broader comparisons or data aggregations will be possible

Implications of Anomalies - 3

- Underdiagnosis- Causes?
  - This was discussed with the Chief Medical Officers at all of the project sites as well as many other staff members (Execs, IT, analysts etc.)
  - Consensus was there are three types of causes:
    - Cultural & Training bias:
      - » Providers, especially doctors, are trained to treat what is in front of them
        - If they are treating a foot infection, they may not even look to see if Type 2 diabetes is on the Problem List or in the clinical data, & they almost certainly will not list it as a diagnosis for the encounter
        - The same is true of most, if not all, comorbidities
    - EHR Characteristics:
      - » Many EHRs have almost separate acute vs. routine encounter workflows
        - If Doctor is treating an acute problem (foot infection), it is difficult to create a record that is complete with respect to co-morbidities
      - » Ease-of-use issues can cause providers to not fill in multiple diagnoses
      - » Lack of workflow alignment can also cause less than optimal use
    - Reimbursement Model
      - » Finally, current reimbursement models do not encourage the reporting or treatment of comorbidities in the same encounter (i.e. CMS PPS model)

Implications of Anomalies - 4

- Underdiagnosis – Implications
  - In at least one case (CHC), comorbidity analysis was not possible as the underdiagnosis ensured that such clusters were inaccurate & not representative of the health centers patient population.
  - CMO estimated 20%-25% of patients were diabetic & 35% of patients were obese, but the only cluster above 2% in the analysis was obesity/ hypertension, CMO's estimate was this should have been at about the 35% level (not 2.4%)
  - If such underdiagnosis is widespread in health center data, which it may be given the consensus on causes, then the whole effort of comorbidity analysis based on diagnosis data in the EHR for cost control & clinical intervention is called into question
    - Also has implications for how providers are trained, how EHRs are designed & how care is reimbursed
    - May be ameliorated by direct analysis of clinical data – we are just starting to try this, but in some cases there are enough errors in the clinical data that the comorbidity analysis – even from clinical data – may be compromised

Some Early Conclusions…

- It **MUST** be emphasized that the results discussed here are **preliminary!**
- Much more work has to be done before any specific or general conclusions can be reached
- Very preliminary interpretation indicates that data quality problems in multi-year EHR at CHCs may be widespread & are the result of technical & usability problems with EHRs, sociocultural & training bias in providers & current reimbursement structure

MIT SOCIOTECHNICAL SYSTEMS RESEARCH CENTER

Thank You

Please feel free to contact me

for more information

David Hartzband

Phone: (212) 246-1122 x 722 (Foundation)

(617) 401-2508 (CIC)

(617) 501-4611 (m)

Email: dhartzband@rchnfoundation.org

dhartz@mit.edu