rchn
community health foundation

If Big Data is the Answer, What's the Question?

Results & Lessons Learned from the P2A Project

David Hartzband, D.Sc.
Director, Technology Research
RCHN Community Health Foundation
&
Research Scholar
Sociotechnical Systems Research Center
Massachusetts Institute of Technology

Path2Analytics (P2A) Project: Goal & Project Structure

Goal

- **The goal of the P2A Project is to introduce contemporary analytics into CHCs so that this capability may be used in the center's strategic decision making**
  - Contemporary analytics simply means the infrastructure, software & expertise to use new analytic capabilities
  - These capabilities are characterized by analysis of very large data sets or data sets with many different types & formats of data
  - This latter is generally more important in healthcare

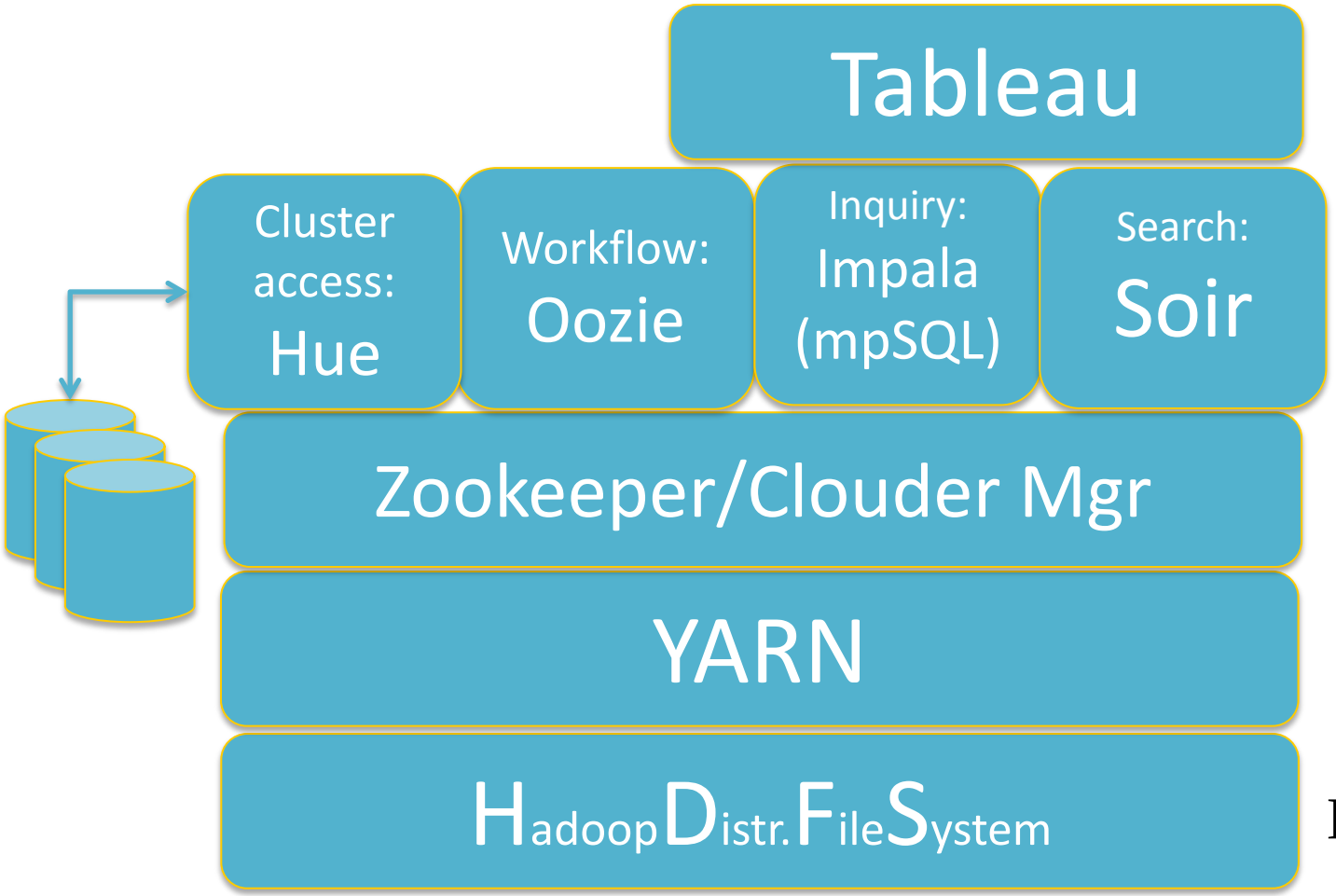
Project Structure

- **This is accomplished by:**
  - Deploying a Hadoop-based, open source analytic stack (Cloudera Express) into the CHC's data center, &
  - Working with the CHC's IT & analytic staff so that they understand & can use the deployment
  - Working with the CHC's Executive Staff so that they understand the difference between the Business Intelligence (BI) & UDS analysis they currently do & the type of analysis made possible by the analytic stack, facilitation strategic inquiry by this staff

Project Status?

- **The Project is currently working with 3 CHCs (2 urban, 1 rural) & a PCA (30 CHCs)**
  - ~400 sites, 1,274,267 patients (2 full data years)
- **Full preliminary results from 1 (urban) CHC**
- **Partial preliminary results from 2 (1 rural, 1 urban) CHCs**
- **Partial preliminary results from the PCA for aggregate & individual CHCs**

Cloudera Express Analytic Stack



Visualization

User Services

Group Services, System Mgmt & Synch

Map Reduce 2

Data Storage/Org

"Big Data"?

- **It is important to emphasize that this is not a "big data" project!**
- **Most healthcare organizations do not have big data, BUT…**
- **Strategic decisions can be supported by relatively small amounts of data, that is the data that a CHC has so long as…**
  - The right questions are asked, &
  - The right data is utilized, &
  - The right analysis is done!

Results & Interpretation

Level-Up Exercise…

- **Purpose is to normalize how data is defined & interpreted between EHR queries & P2A queries (done on analytic stack)**
- **Determine # patients/year**
- **Determine #patients/diagnosis/year**
  - Hypertension, diabetes, obesity, heart disease, behavioral
    - UDS definitions for patients, visits, conditions
- **Determine most prevalent comorbidities/population**
- **Determine cost & revenue associated with:**
  - Condition/patient, condition/location…
  - Comorbidity/patient, comorbidity/location

CHC1 – Urban: Condition Percentages

	CHC1 %ages		P2A %ages		U.S.cuml.%
Dx	2013	2014	2013	2014	
Hypertension	17%	16%	19%	18%	29%
Diabetes	8%	8%	8%	8%	9%
Obesity	**9%**	**10%**	**9%**	**9%**	35%
Heart Disease	**1%**	**1%**	**3%**	**3%**	11%
Behavioral	19%	21%	17%		25%
	21067	23621			

- Results have been normalized to within ~2%
- 2012 data found to be incomplete

CHC2 – Rural: Condition Percentages

	CHC2 %			P2A %			U.S. %
	2012	2013	2014	2012	2013	2014	
Hypertension	16%		14%	4%	4%	4%	29%
Diabetes	6%		6%	2%	2%	2%	9%
Obesity	3%		2%	2%	2%	2%	35%
Heart Disease	4%		3%	3%	2%	2%	11%
Behavioral	*						25%
Patients	15939	17064	17918	17330	18058	18818	

* Data not available

- Results not normalized yet, *i.e.* we do not know why the figures from the EHR queries are so much higher than the numbers from the analytic stack queries

PCA: Condition Percentages - Aggregate

	PCA %			P2A %			U.S.%
DX	2012	2013	2014	2012	2013	2014	
Hypertension	20.73	22.46	22.78	20.5	23.04	23.24	29%
Diabetes	6.59	6.56	6.58	6.49	7.26	6.70	9%
Obesity	4.38	10.00	11.68	4.37	10.23	11.78	35%
Heart Disease	1.43	1.59	1.63	1.39	1.61	1.63	11%
total Patients	350311	403286	440713	331968	384515	421159	

- Results aggregated from 30 CHCs
- Data for P2A analysis loaded from data extracts, already normalized
- Results needed to be adjusted for UDS definitions

Preliminary Interpretation

- **Initially all CHCs were using non-UDS definitions for patient, visit, etc.**
  - Needed to enforce use of UDS definitions in order to create standardized data
- **Differences in numbers (patients, diagnoses/patient/year, etc.) needed to be evaluated at the level of the SQL queries**
  - Often this was the only way to determine what definition was being used
  - In at least one case, the SQL could not be evaluated because the CHC used an intermediate BI tool to do queries that did not make the actual query visible
- **Percentages for all diagnoses lower than expected - much lower than the CDC figures for the U.S. population - but Obesity & Heart Disease very low**
  - Some sociocultural issues suggested for low obesity figures, heart disease figures not currently understood

Summary and Lessons Learned

Summary

- **DEFINITIONS MATTER!**
  - Agree first on definitions, then do analysis…
  - Level-up exercise (matching definitions) took 6-7 months at one urban CHC & is still underway at one urban & one rural CHC
  - The PCA had already done normalization for their data warehouse & also used the UDS definitions as recommended, so level-up exercise took 2 weeks
- **Most diagnosis percentages (patient/diagnosis/year) were at or below CDC figures for U.S. population**
  - Does not meet our expectation for CHC populations
- **Obesity & especially Heart Disease percentages were very low (Obesity <10%, HD <5%)**
  - Causes under investigation
- **Comorbidity percentages not usable because condition percentages are so low**

EHR Structure & Function

- **In at least one case, the EHR in use treated ICD-9 expressions differently on query**
  - 250, 250.0, 250.00, 250* produced different results, as did 250.5 & 250.50
- **Vendors did not want to make underlying DB schema visible**
  - In one case, schema was so complex that what tables were being queried was not possible to determine (>1000 tables)
- **Navigation for anything but the simplest workflows is difficult enough to discourage use**
  - Acute workflows not conducive to treatment of multiple conditions
- **Little support for multiple diagnoses/encounter**
  - Use case informally tested at HIMSS this year with 6 EHR vendors – none provided adequate support
- **Many deleterious effects of migration from one EHR to another & unsuccessful integration in general**

Sociocultural Effects

- **Low obesity diagnoses discussed with many CMOs & CHC staff**
  - Most agree that this is not surprising, although most CMOs thought their population was in the range of 40-45%
  - Many CMOs say their providers do not use the full range of UDS specified ICD-9 codes (V-codes for specification of overweight, obese etc.)
  - Many CMOs, providers & staff members believe that there is a social bias against making this diagnosis
- **Heart Disease percentages (<5% when the U.S. population mean is 11%) are harder to interpret**
  - Most CMOs felt that their population might be in the range of 20%-30%
  - Cause still under investigation

Big Data & Big Lessons…

- **"Big" data not necessary to do analytics!**
  - Many strategic questions can be addressed with the data at hand
    - Each CHC has 5-10 GBs of EHR & up to 5 GBs of other data (financial, cost accounting, registry, pharmacy etc.)
    - PCA has ~250GBs of EHR data & 100GBs of other data (as above)
    - Each expected to double in the next 2-3 years plus need to include an additional 20-25GBs of external data (public health, macrodemographic (State & Federal), macroeconomic (State & Federal) data
  - Largest healthcare organizations already have ~50 PBs (1024 TBs, 10^{15} bytes) of data (Kaiser)
- **Asking the right questions is more important than having the most data!**

Big Data & Big Lessons 2…

- **How do we ask the "right" questions?**
  - Being "data aware" allows us to focus on strategy rather than analysis
  - Need to know:
    - What data is available (not just EHR, PM)
    - Quality of data & how that may limit use
    - Types of analysis available (including modeling & prediction)
- **Involving people from multiple parts of the CHC is important in developing analyses: Executive, Admin, Clinical all must be represented**
- **Essential to understand the culture of providing care in developing questions & interpreting analytic results (*e.g.* see Obesity results)**

Almost There…

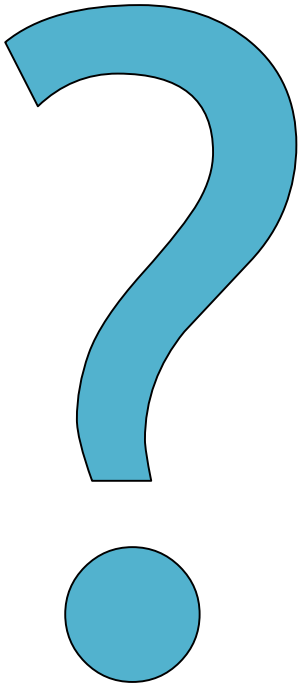
- Just doing "level-up" exercise & initial strategic analysis will bring up many issues to be addressed as well as strategic results that can not be developed in other ways
- The process of getting to be able to do analytics is initially as important as doing them
- Spreading the data net as widely as possible allows for more compelling results
- Getting to analytics will allow a CHC to be positioned for the next steps in the evolution of information management & use
- "Just do it…", but why???

Change is Good, Right?

- **The information management environment for CHCs (& healthcare organizations in general) is changing**
- **In the next 3-5 years:**
  - CHCs will have double to triple the amount of data they currently use including considerable external data from many sources
  - New organizational models (HIEs, ACOs…) new regulatory requirements (Meaningful Use 3) will require CHCs to make much more use of all of this data to be able to be able to provide care & operate in the new environment
  - Technology is changing quickly & current best practices for information storage, management analysis & usage will be obsolete shortly (2-3 years)
- **Beginning the exploration of analytics will position the health center to be able to evolve with the environment**

& Finally,… The Question

- **The real question is "if analytics is the answer, what's the question?**
- **The question is:**
  - What do you do in order to position your CHC for the next steps in the evolution of information management & use for strategic decision-making?
- **& the answer is**
  - **Analytics**



Thank You

Please feel free to contact us
for more information

David Hartzband

RCHN Community Health Foundation

55 Broadway, Suite 1502

New York, New York 10019

Phone: 617.501.4611

Email: dhartzband@rchnfoundation.org