More Than A Buzzword:
Big Data in the Environmental Arena

2015 National Environmental Monitoring Conference | July 15, 2015

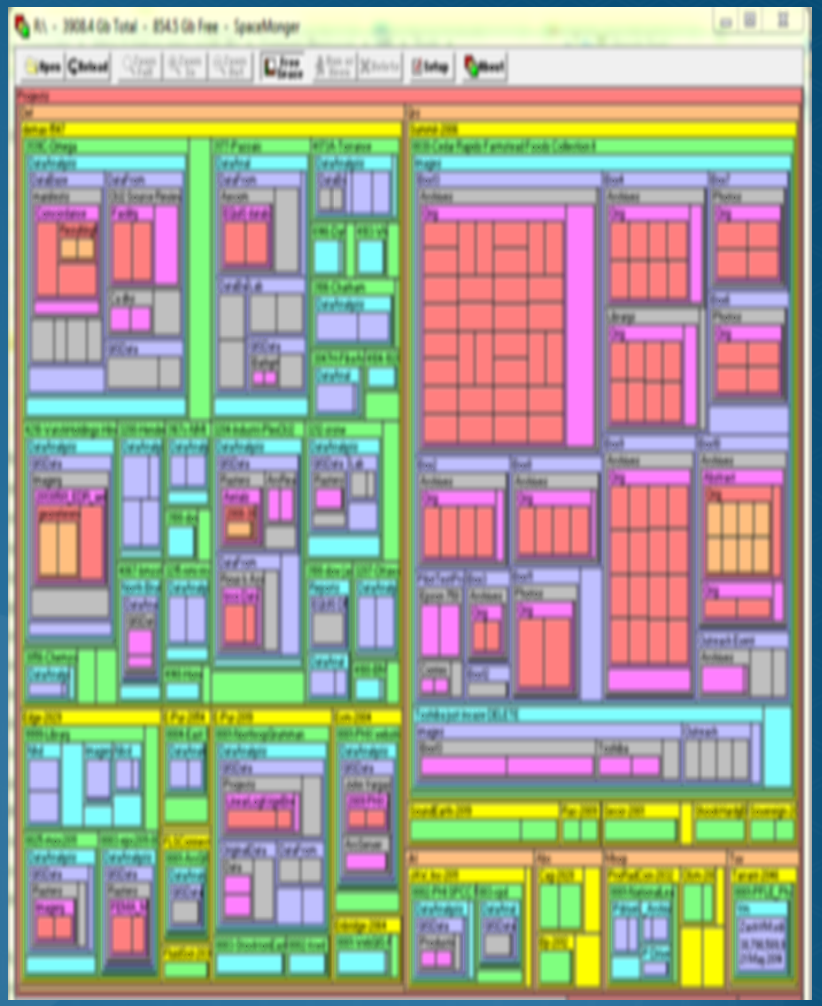
Brooke Roecker
Senior Environmental Data Analyst

Mark Packard, PG, CPG
President/CEO

ddms™
www.ddmsinc.com

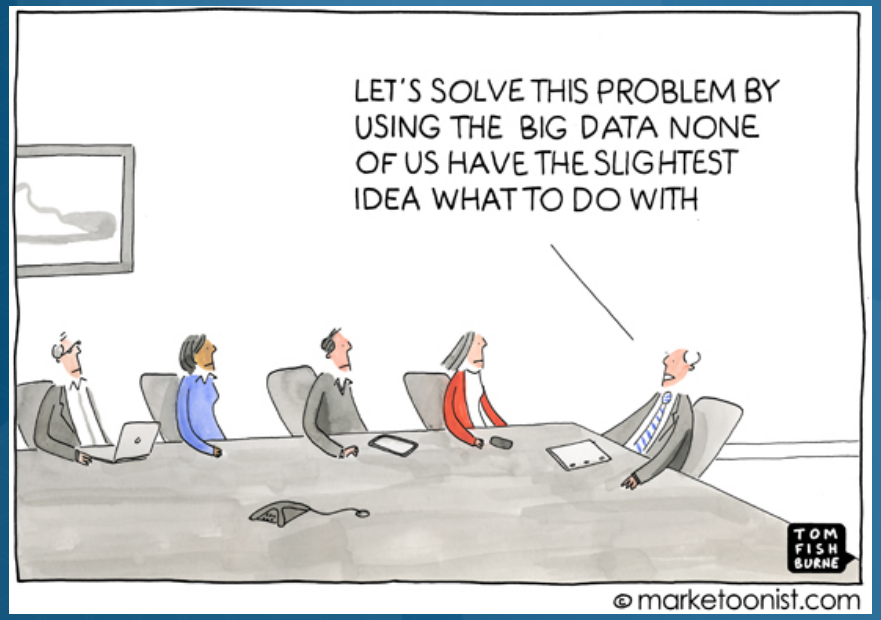
Presentation Outline

- Big Data Defined
- Environmental Data Past & Present
- Today's Tools and Approaches
- Example Projects
- Future Considerations



ddms

Big Data in Environmental Monitoring/Remediation



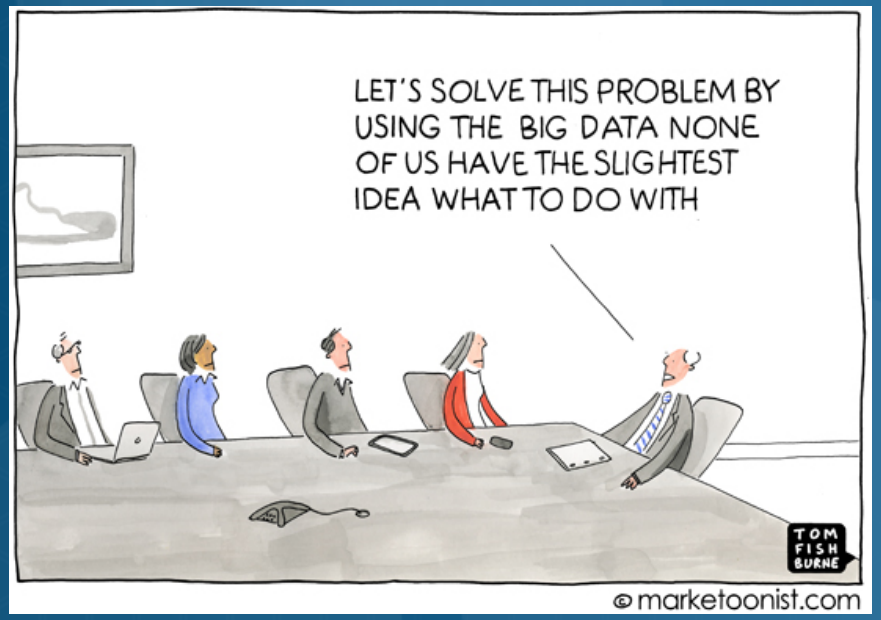
Big Data Defined: (yet again)

"Big data is high volume, high velocity, and/or high variety information assets that require new forms of processing to enable enhanced decision making, insight discovery and process optimization."

Laney, Douglas. "The Importance of 'Big Data': A Definition". Gartner. Retrieved 21 June 2012.

ddms

Big Data in Environmental Monitoring/Remediation



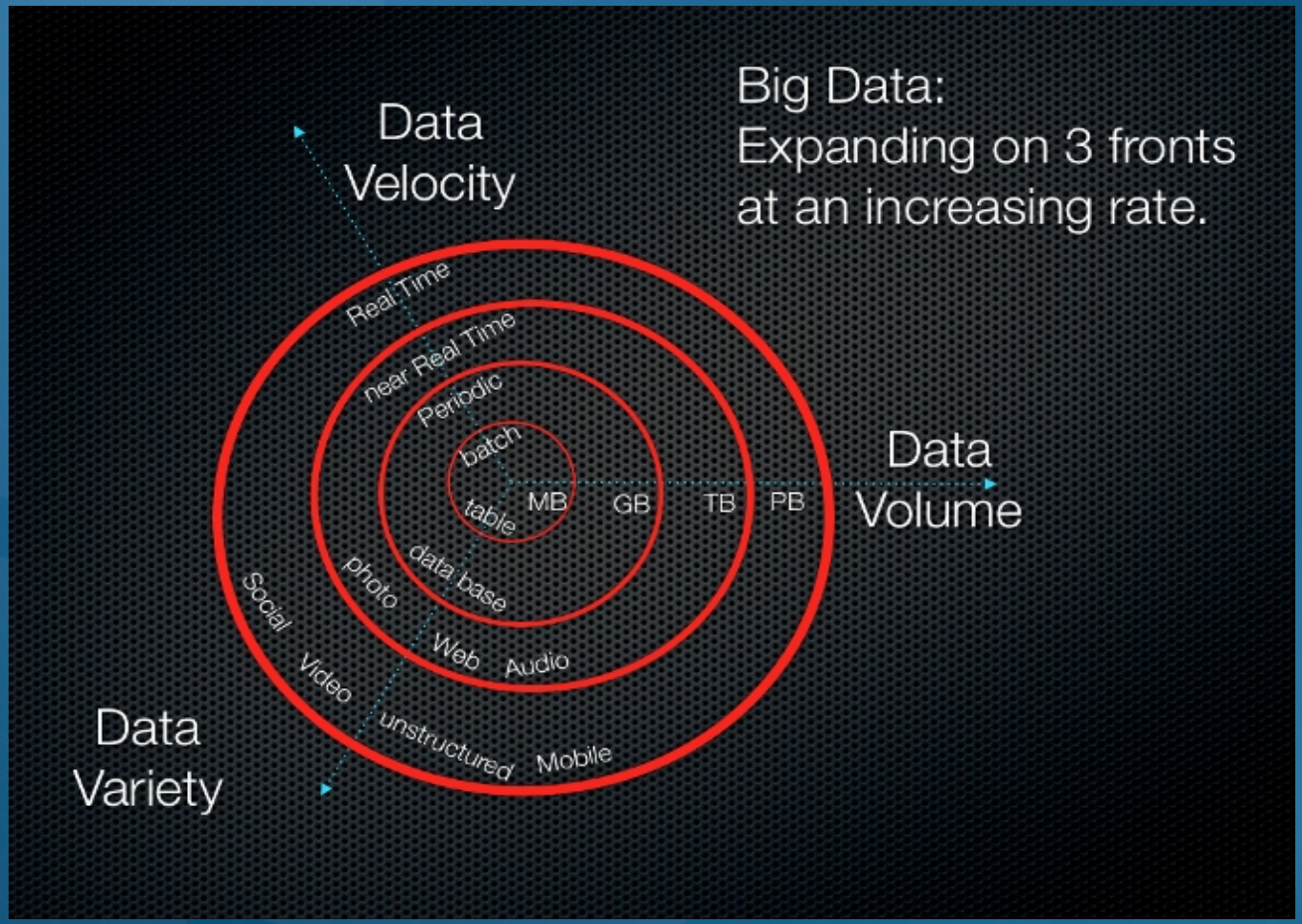
Big Data Defined: (yet again)

"Big data is high volume, high velocity, and/or high variety information assets that require new forms of processing to enable enhanced decision making, insight discovery and process optimization."

Laney, Douglas. "The Importance of 'Big Data': A Definition". Gartner. Retrieved 21 June 2012.

ddms

- Velocity: high frequency data
- Variety: mixed data/attributes
- Volume: very large datasets
- VERACITY: Accuracy of data



Diya Soubra, "The 3Vs that define Big Data". http://www.datasciencecentral.com/forum/topics/the-3vs-that-define-big-data

ddms

Where have we come from?

- Hand-written field logs
- Text files
- Spreadsheets
- Simple Reports

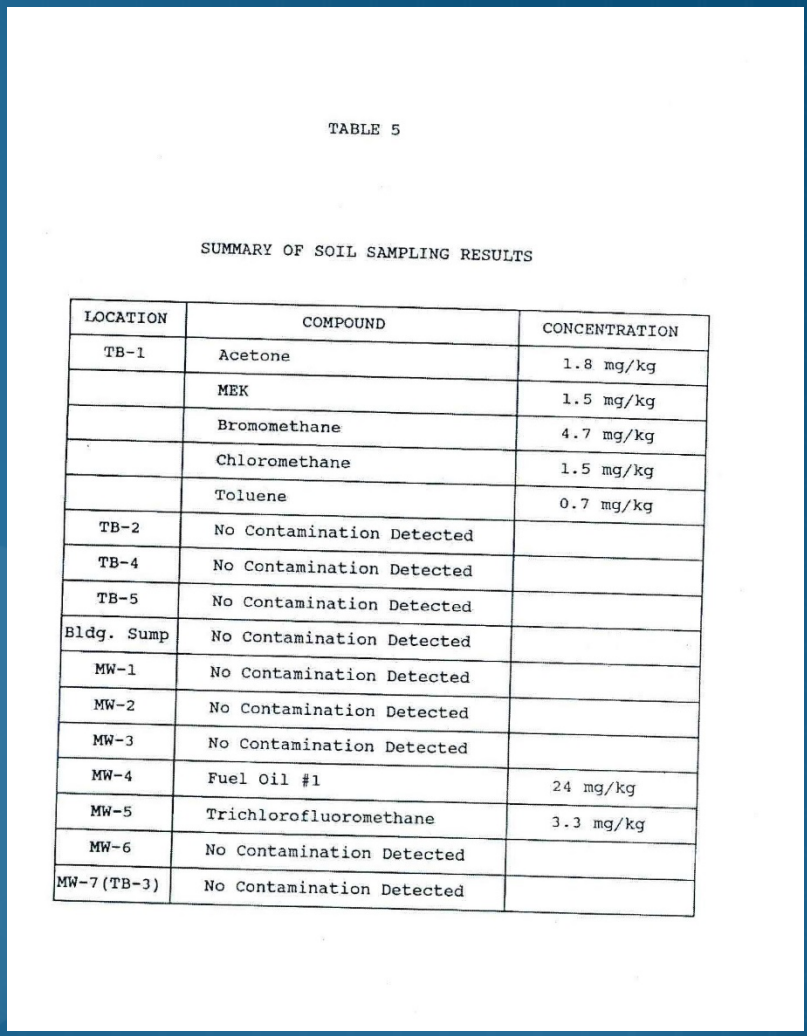


TABLE 5

SUMMARY OF SOIL SAMPLING RESULTS

LOCATION	COMPOUND	CONCENTRATION
TB-1	Acetone	1.8 mg/kg
	MEK	1.5 mg/kg
	Bromomethane	4.7 mg/kg
	Chloromethane	1.5 mg/kg
	Toluene	0.7 mg/kg
TB-2	No Contamination Detected	
TB-4	No Contamination Detected	
TB-5	No Contamination Detected	
Bldg. Sump	No Contamination Detected	
MW-1	No Contamination Detected	
MW-2	No Contamination Detected	
MW-3	No Contamination Detected	
MW-4	Fuel Oil #1	24 mg/kg
MW-5	Trichlorofluoromethane	3.3 mg/kg
MW-6	No Contamination Detected	
MW-7(TB-3)	No Contamination Detected	

Where are we today?

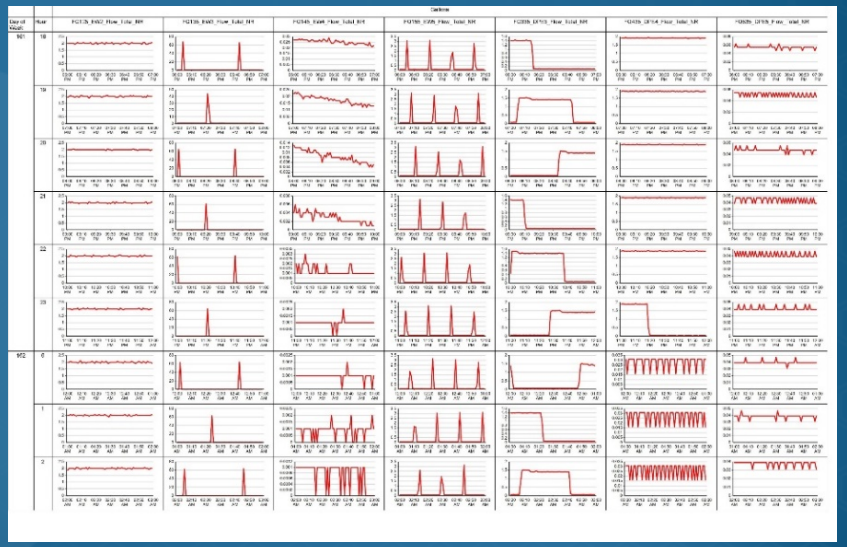
- Older technologies remain
- Database storage
- Out of the box storage/analysis tools
  - EQuIS™
  - ENFOS
  - Locus
  - Project Portal™

Where are we today?

- Older technologies remain
- Database storage
- Out of the box storage/analysis tools
  - EQuIS™
  - ENFOS
  - Locus
  - Project Portal™
- Limitations?
- What data are you not managing/analyzing?

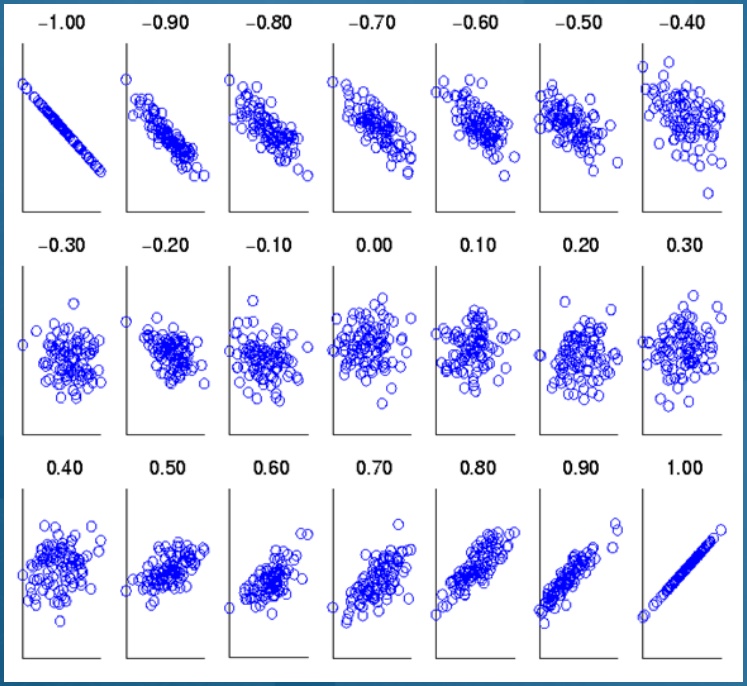
“Bigger” data tools available today

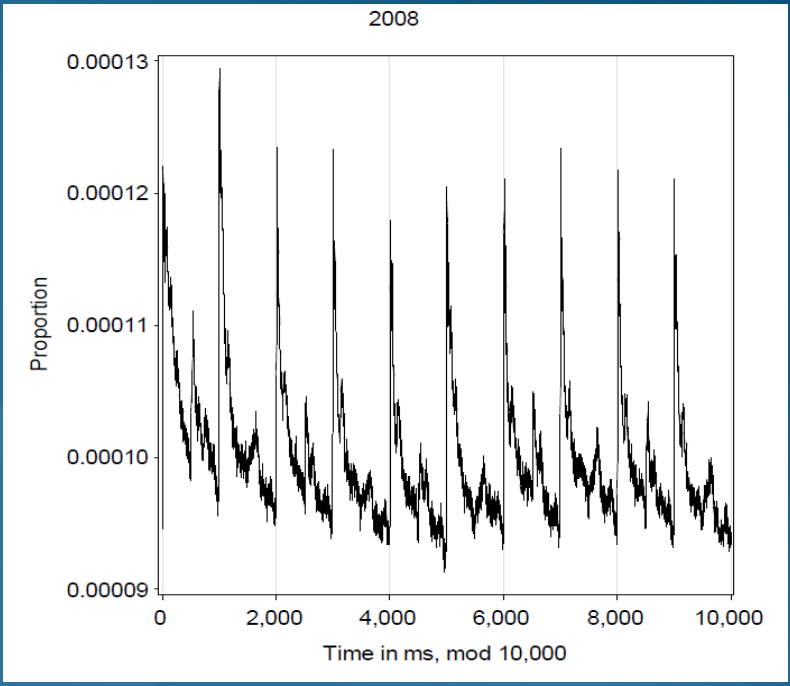
- High frequency advancements:
  - EQuIS Live
  - Project Portal Analytics Module
- Analysis and modeling tools
  - Spatial: ArcGIS, EVS
  - Visualization: Tableau
- Custom scripting (R, Python, T-SQL…)
  - SSAS, Weka
  - MatPlotLib (Python), ggplot2 (R)



ddms

Project Examples





ddms

Project: Surface Water Monitoring

- High velocity data, 5-minute intervals
- Teledyne ISCO samplers
- Historical data archived in raw MS Excel files

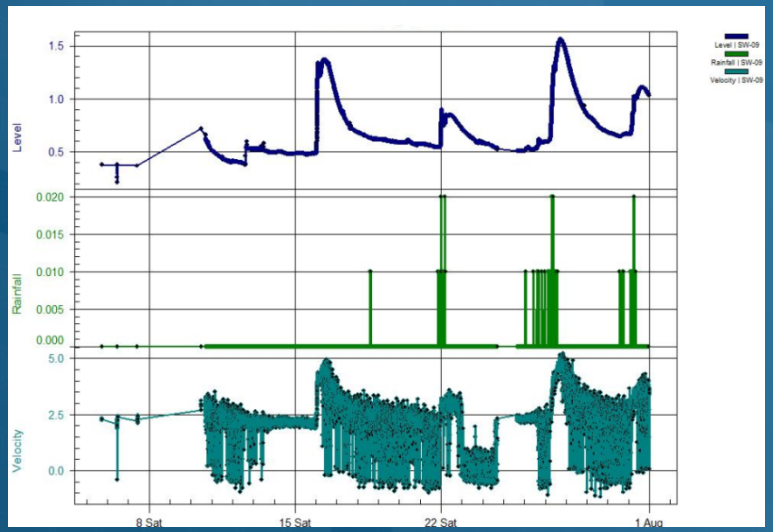
Challenge:

- Dataset too large for “Big picture” trends
- Storing/archiving data long-term
- Centralized access for project team

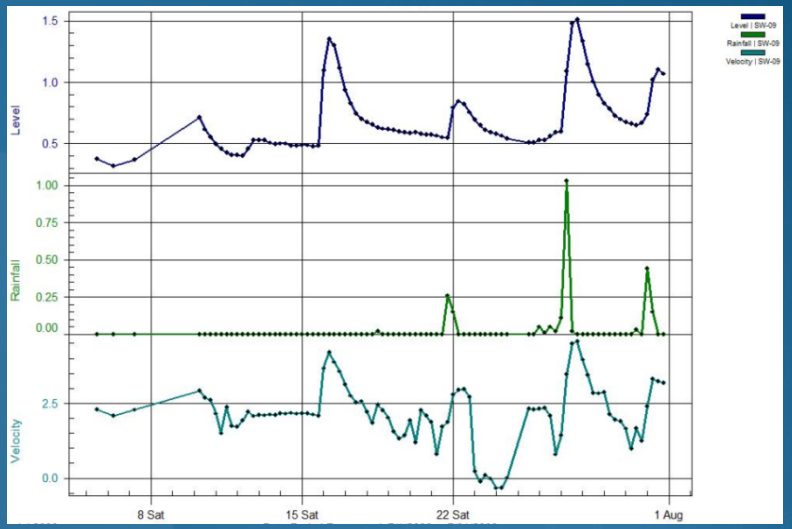
Project: Surface Water Monitoring

Solution:

- Streamlined data resampling & import routine
  - Resample from 5-minute to 12-hour averages (or totals)



Raw Data



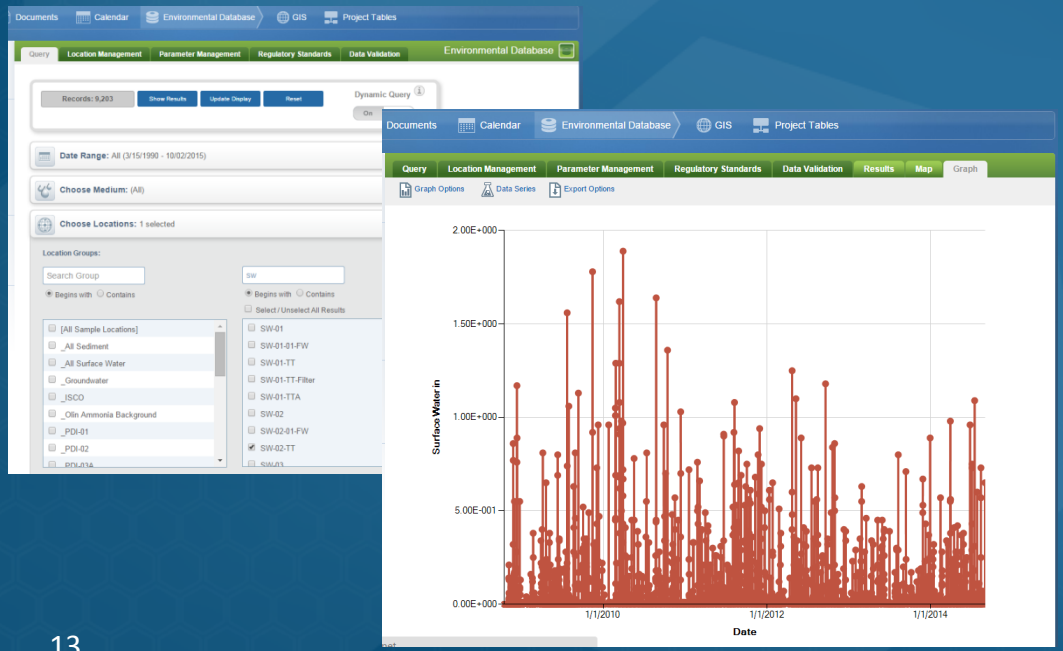
Resampled Data

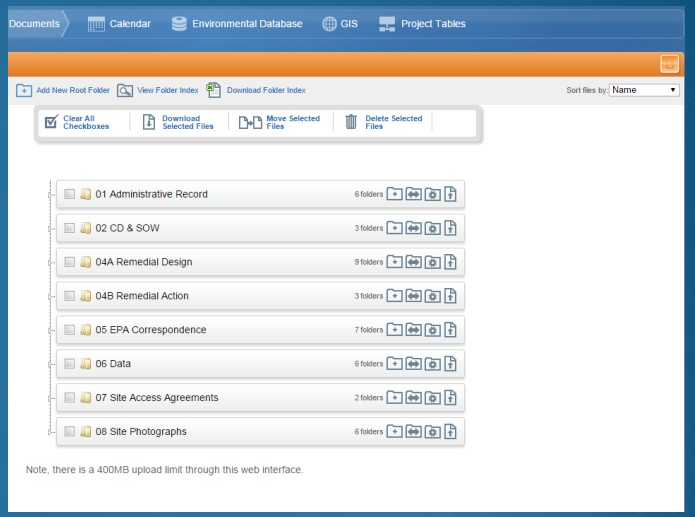
ddms

Project: Surface Water Monitoring

Solution:

- Data available to project team via Project Portal
  - Environmental Database module for resampled data
  - Documents module for raw data



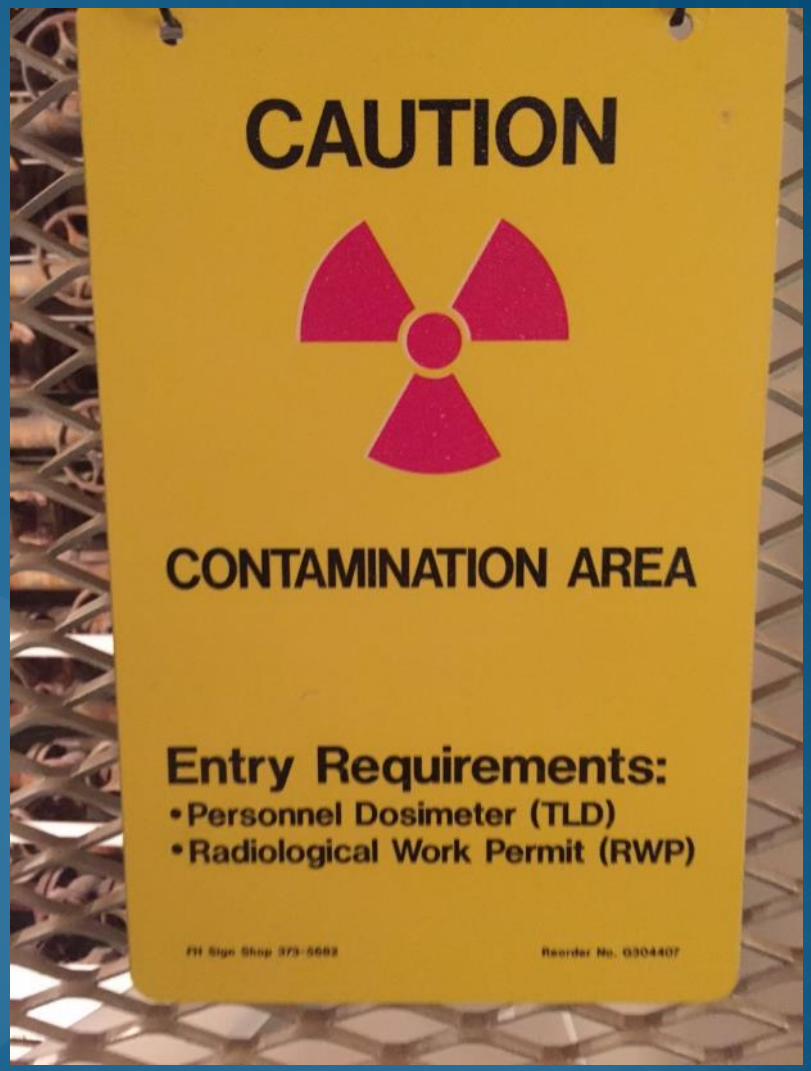


Project: RAD Site Monitoring

Challenge:

- High volume, high velocity sensor data w/ telemetry
- Automated database storage
- Visual analysis of high volume weather data

Date	Time	Baro	EPM1	EPM2	EPM3	EPM4	Prcp	RH	Temp	WDir	WSpd	Pwr
3/4/2015	16:07:03	29.818	8	11	2	13	0.23	45.807	39.635	291.894	9.322	1
3/4/2015	16:09:20	29.818	8	9	2	10	0.23	45.984	40.434	253.336	6.118	1
3/4/2015	16:10:37	29.819	7	7	2	11	0.23	45.415	37.941	262.806	3.628	1
3/4/2015	16:13:55	29.821	7	8	2	10	0.23	44.88	38.819	222.352	4.42	1
3/4/2015	16:15:13	29.819	7	-1	2	11	0.23	44.52	38.819	269.378	6.369	1
3/4/2015	16:16:29	29.818	6	7	2	13	0.23	44.371	39.625	285.836	4.844	1
3/4/2015	16:19:47	29.814	6	5	3	7	0.23	43.91	40.218	260.142	3.879	1
3/4/2015	16:21:04	29.817	5	10	4	8	0.23	43.901	37.768	255.619	6.041	1
3/4/2015	16:22:22	29.818	4	13	2	8	0.23	45.294	36.176	291.813	4.188	1
3/4/2015	16:24:39	29.82	5	12	3	7	0.23	47.733	38.514	253.603	8.01	1
3/4/2015	16:25:57	29.821	4	12	2	6	0.23	47.713	39.113	265.432	7.315	1
3/4/2015	16:27:15	29.819	3	11	2	7	0.23	48.261	36.859	279.921	7.99	1
3/4/2015	16:28:32	29.819	2	10	2	8	0.23	47.978	37.012	255.555	4.883	1
3/4/2015	16:29:49	29.818	3	10	2	8	0.23	47.732	37.012	268.397	5.636	1
3/4/2015	16:32:06	29.819	2	9	2	9	0.23	47.142	37.127	246.404	4.844	1
3/4/2015	16:33:23	29.819	4	8	2	7	0.23	47.142	37.127	258.235	5.192	1
3/4/2015	16:34:41	29.819	5	7	2	8	0.23	46.491	38.128	234.716	4.825	1
3/4/2015	16:38:58	29.823	8	6	2	8	0.23	48.924	35.301	251.514	6.755	1
3/4/2015	16:40:16	29.824	8	6	2	8	0.23	48.573	37.297	246.073	6.427	1
3/4/2015	16:41:33	29.825	5	5	2	7	0.23	48.314	38.09	254.202	8.087	1
3/4/2015	16:44:51	29.827	5	4	2	8	0.23	48.959	37.012	245.586	5.925	1
3/4/2015	16:46:08	29.828	6	3	2	7	0.23	48.172	35.853	260.303	4.864	1
3/4/2015	16:47:26	29.828	6	3	6	9	0.23	47.928	36.646	255.07	3.493	1
3/4/2015	16:48:43	29.829	5	3	11	13	0.23	47.715	37.244	264.796	5.057	1

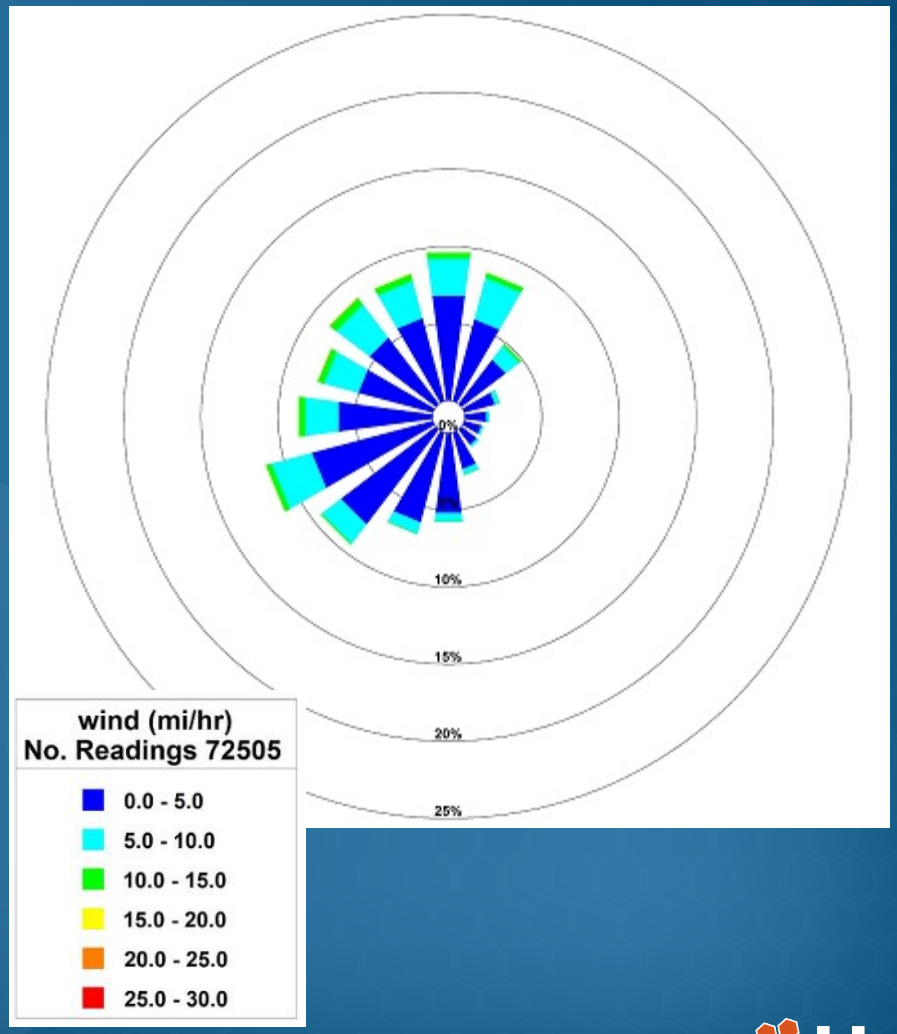


ddms

Project: RAD Site Monitoring

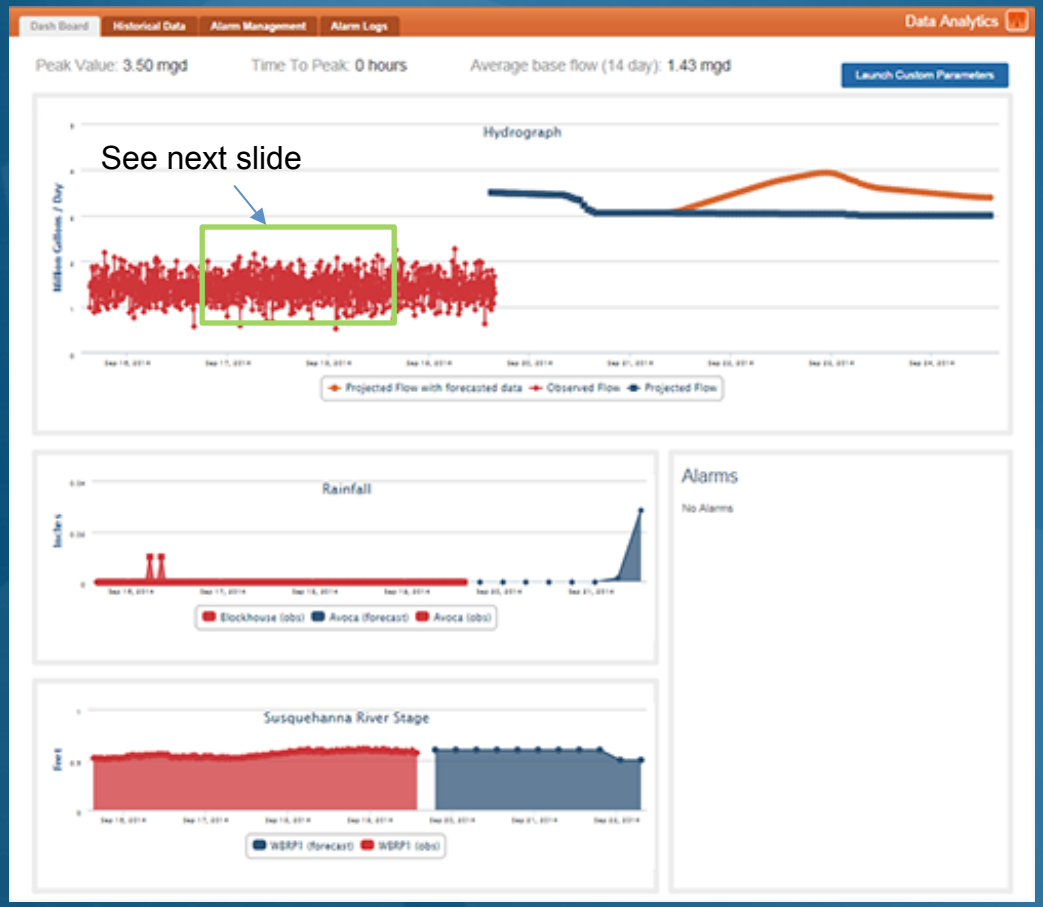
Solution:

- Automated data import via web upload
- Database available to field staff via Project Portal
- Wind rose graphics to visualize data via EnviroInsite
- QA of erroneous data points (particulates)



ddms

Project: Mine Tunnel Monitoring



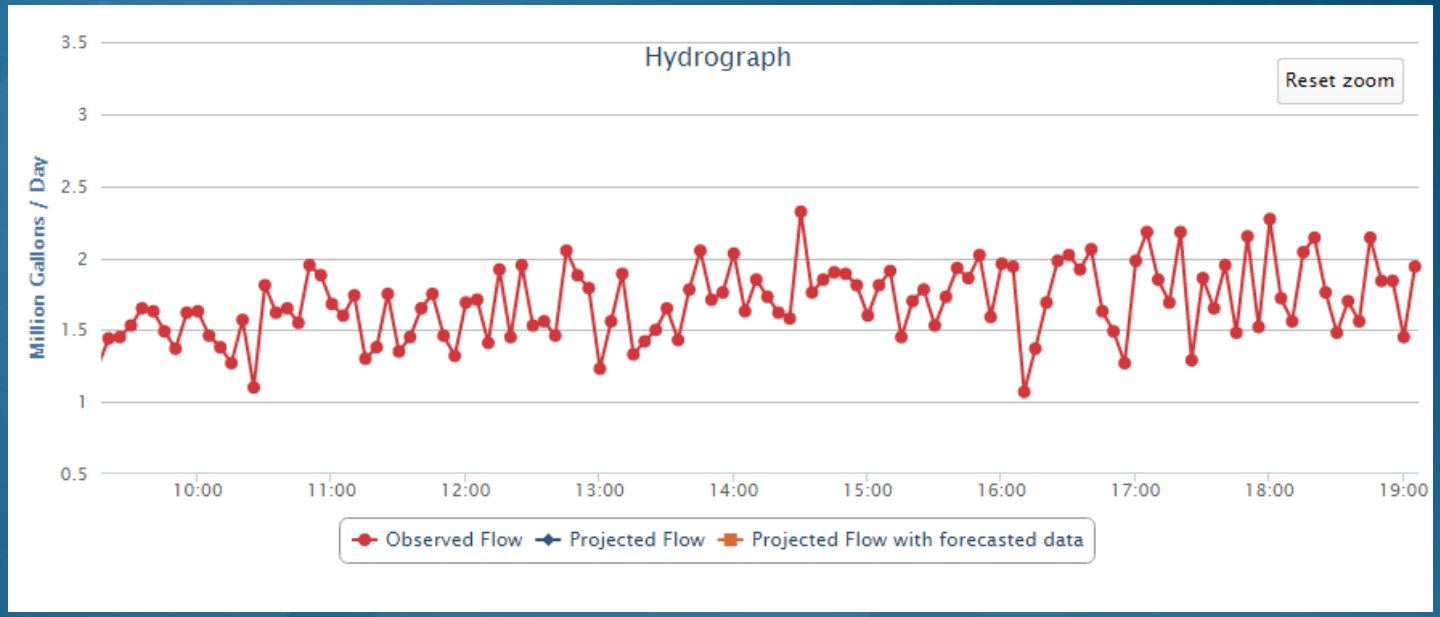
- High volume, high velocity data
  - On-site sensor data from PLC system
  - Public big data streams

Challenge:

- Centralized database storage
- Real-time data access
- Real-time notifications/ alarms

ddms

Project: Mine Tunnel Monitoring

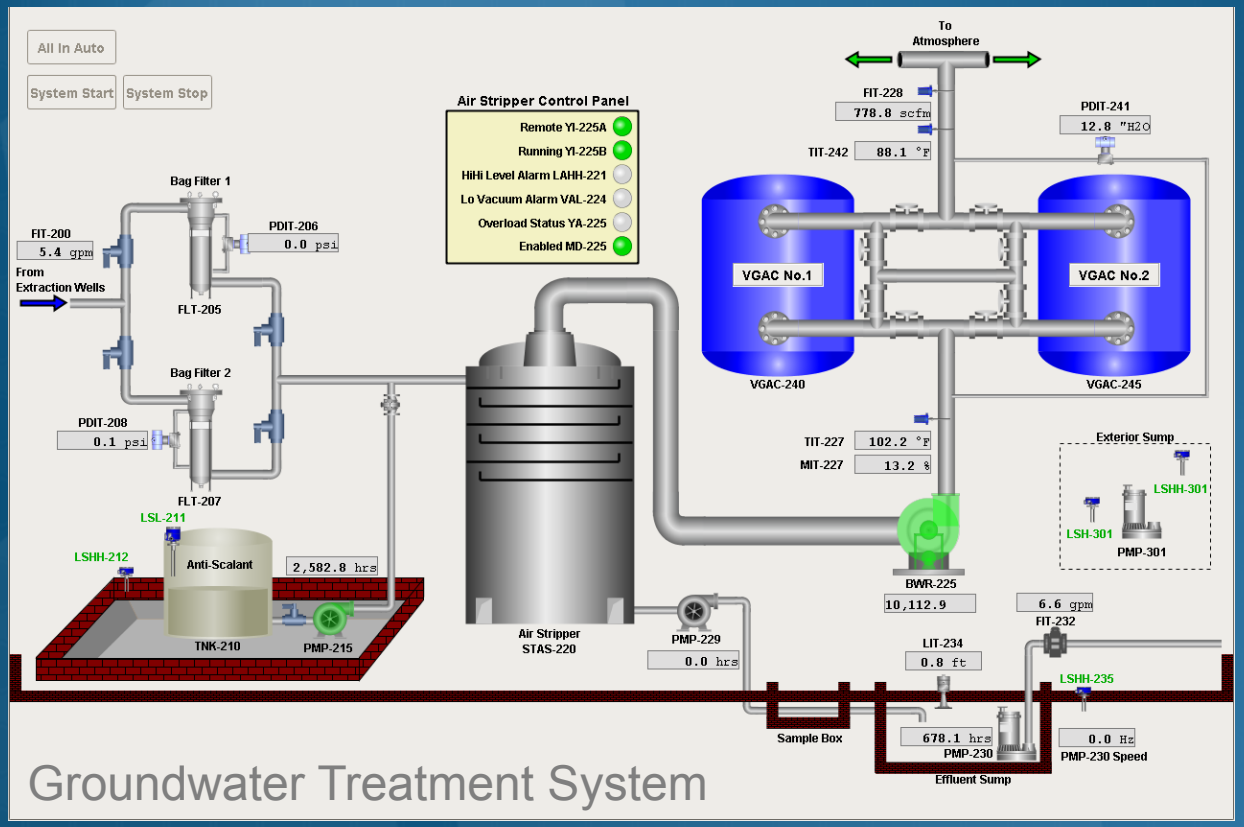


Solution:

- Generic database design optimized for high frequency data
- Predictive trend modeling calculations
- Data available via Project Portal
- Email alerts when incoming data parameters out of spec.

ddms

Project: O&M Site Monitoring

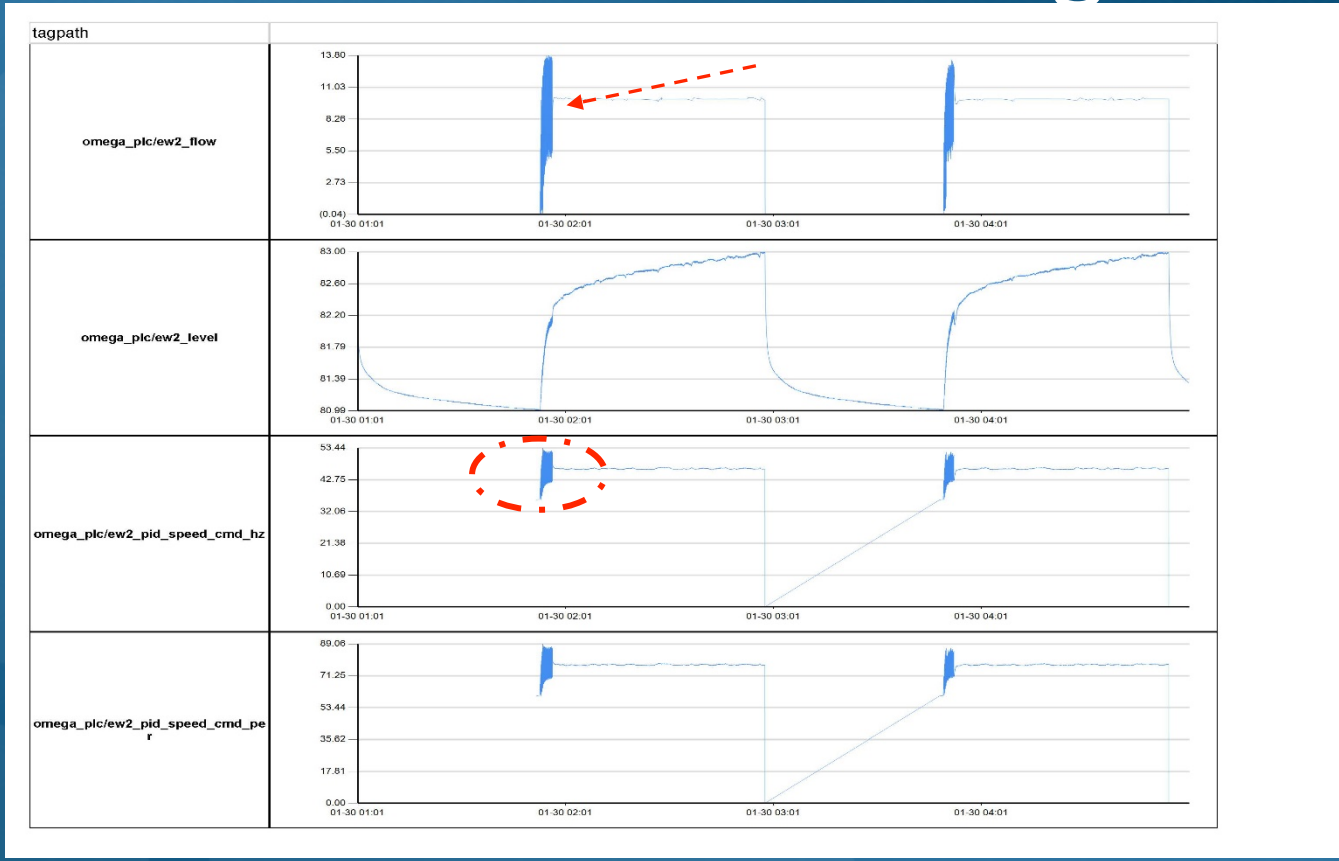


High velocity, automated SVE and GW treatment systems

Challenge:

- Centralized storage
- Centralized monitoring
- System troubleshooting…

Project: O&M Site Monitoring



System Solution:

- Multivariate analysis to review system variables
- Secondary analysis to identify fluctuations

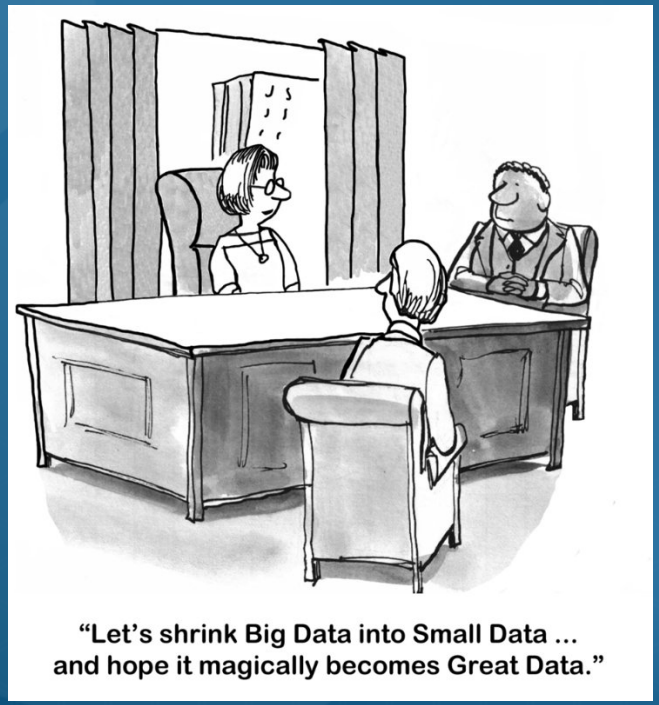
ddms

What's Next?

- Use of emerging technologies:
  - Distributed data sourcing
    - Hadoop HDFS
    - NoSQL
  - Distributed processing
    - Batch processing (MapReduce, Apache Hive)
    - Real-time processing/streaming (Cloudera Impala, Apache 54)
    - PolyBase (cross-querying HDFS and SQL Server)

Summary / Key Takeaways

- Free big data – go out and use it!
- Big data to investigate the unknown
- Greater project intelligence & decision making



linkedin.com/pulse/auto-industry-bored-big-data-brian-pasch

ddms

Thank you!

Contributors:

Myles Hook, ddms
Jon Turner, ddms
Heidi Gaedy, ddms
Ed Larson, ddms
Angela Remer, ddms
Emily Mulford, Earthsoft Inc.

Brooke Roecker
BRoecker@ddmsinc.com

Mark Packard, PG, CPG
MPackard@ddmsinc.com

ddms™
www.ddmsinc.com