Technische Universität München

sebis

Master's Thesis: Descriptive study and experimental analysis of the ELK stack applicability for Big Data use cases

March 21th, 2016

Oemer Uludag, Prof. Dr. Matthes, Thomas Reschenhofer, M.Sc.

Software Engineering for Business Information Systems (sebis)
Department of Informatics
Technische Universität München, Germany

wwwmatthes.in.tum.de

Agenda

1. Introduction
2. Related work
3. Descriptive study
4. Case study
5. Cross-case analysis
6. Conclusion and outlook
7. Near real-time Twitter analysis

Introduction: Motivation

Technological transformation in the area of mobility

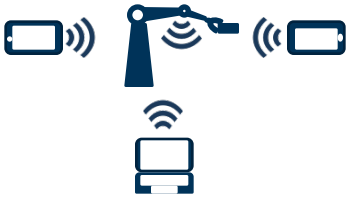
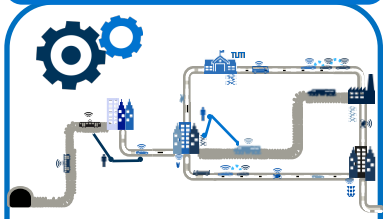
- Ubiquitous computing is also in the area of mobility, promoting to new technologies and leading to a rapid and disruptive technological transformation in this area.
- Various kinds of vehicular sensors generated by the Internet of Things and a new generation of strongly networked and integrated systems contribute continuously to the expansion of huge mounds of data.
- The ability to process and analyze this data and to extract insight and knowledge that enable intelligent services is a critical capability.

Source: based on [1, 2, 3]

Introduction: Motivation

Technological transformation in the area of mobility

Example of these kinds of applications in the mobility industry comprise:

Connected vehicle	Autonomous driving	Smart manufacturing	Industrial internet	Mobility services
				

The 5 Vs of volume, velocity, variety, veracity, and value are often used to describe the requirements of Big Data applications and the characteristics of Big Data.

Source: based on [3]

Introduction: The ELK stack

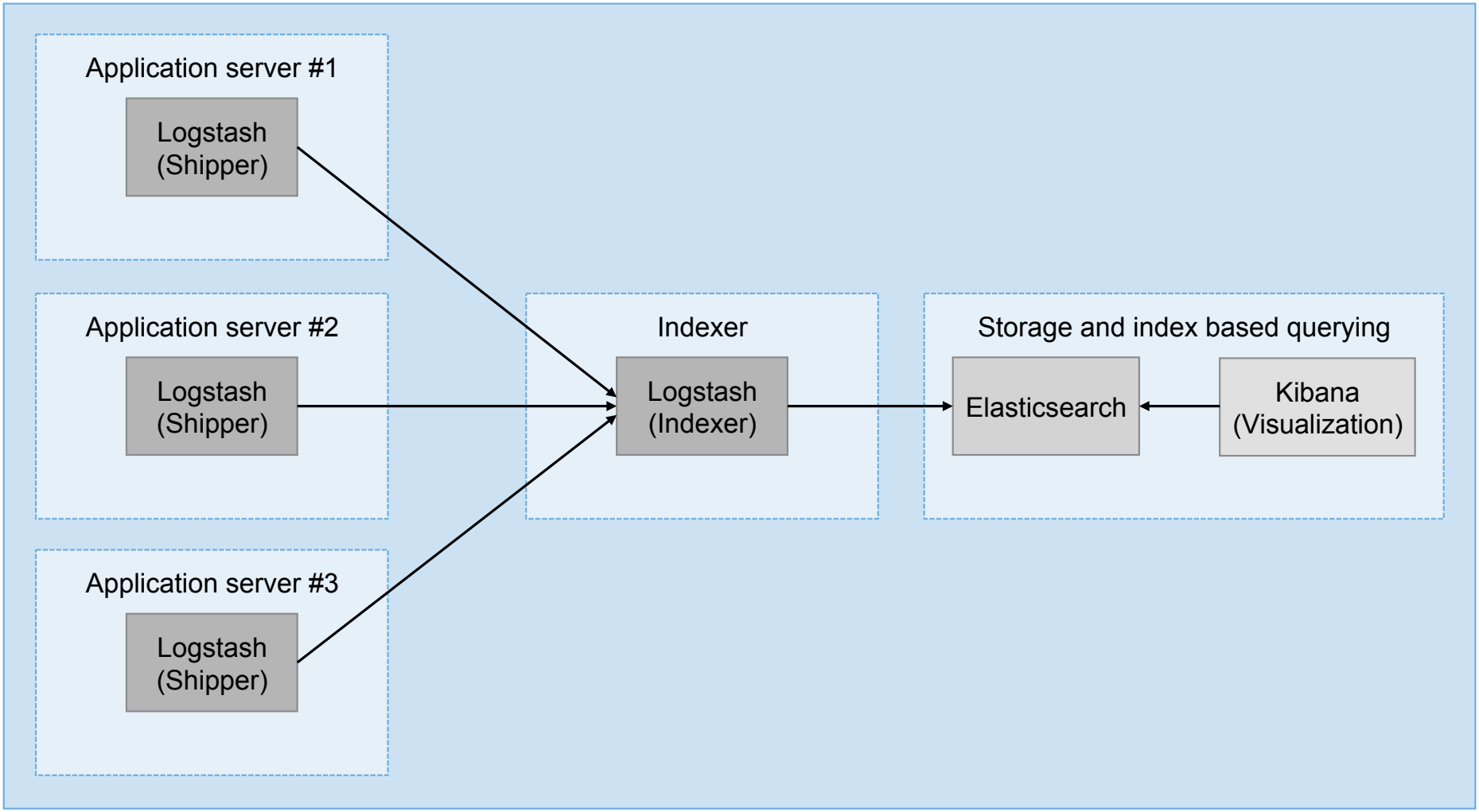
The Elasticsearch, Logstash and Kibana (ELK) stack as an outstanding search-based data discovery tool

- The ELK stack is and end-to-end stack that glean actionable insights in near real-time from almost any type of structured and unstructured data source.
- **Elasticsearch**: performs deep search and data analytics.
- **Logstash**: is responsible for centralized logging, log enrichment and parsing log files.
- **Kibana**: is used to visualize data from Elasticsearch.

Due to the fact that the ELK stack is used by many organizations for a variety of business critical functions, an evaluation of its applicability in the mobility industry seems necessary.

Source: based on [4, 5]

Introduction: The ELK stack architecture



Source: based on [6]

Introduction: Research questions

Research question 1:

What are capabilities and key features of the ELK stack?

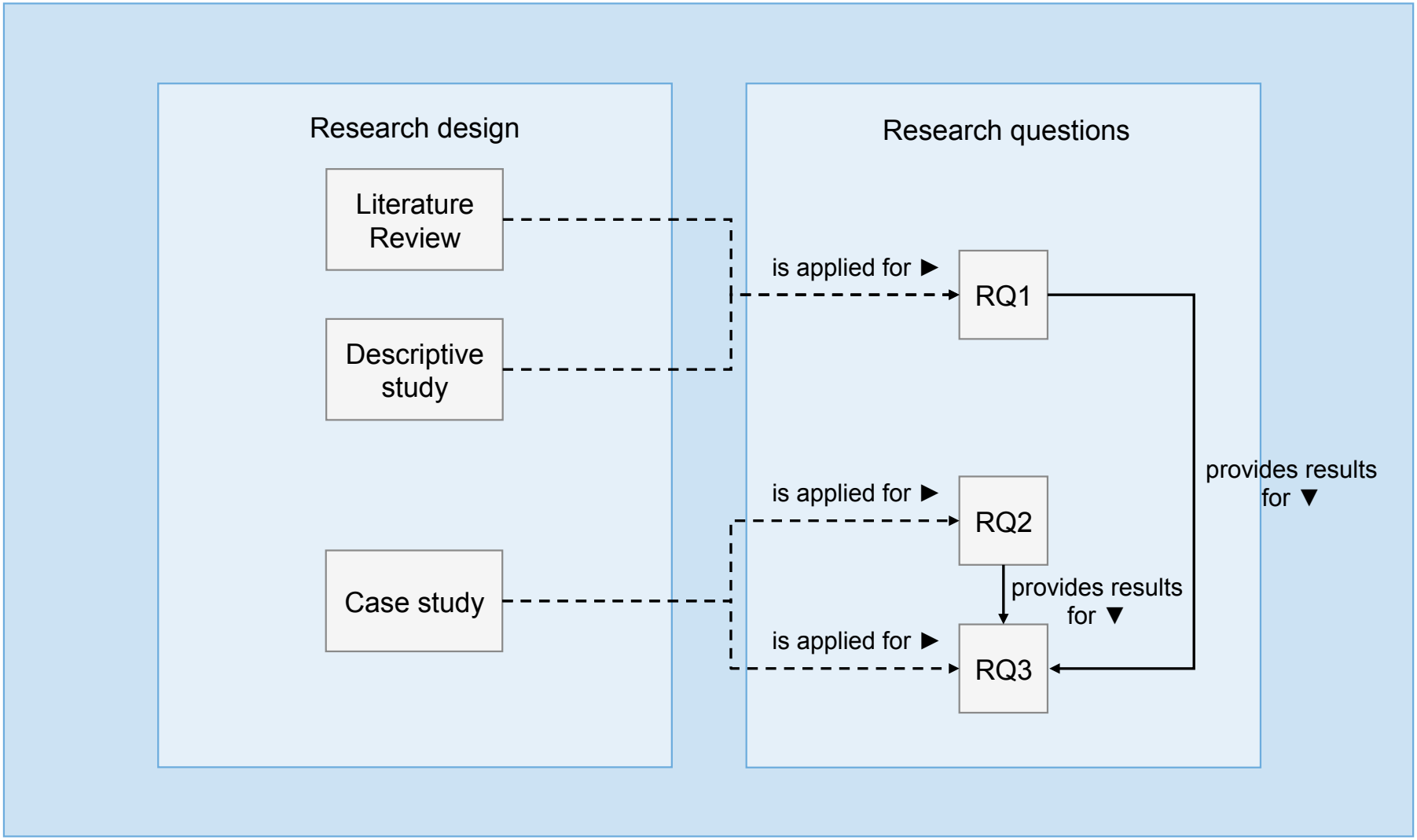
Research question 2:

What are Big Data use cases in the mobility industry?

Research question 3:

For which type of Big Data uses cases is the ELK stack applicable?

Introduction: Research approach



Introduction: Contribution

The contributions of this master's thesis include

- Provision of an holistic view of the main characteristics and functionalities of the ELK stack
- Comparison of Elasticsearch and ELK stack to other related technologies
- Description of Big Data use cases in the mobility industry
- Compilation of Big Data use case requirements
- Juxtaposition of Big Data use cases on the architecture and the technology of the ELK stacks using requirements

Related work: Key findings of literature review

Mention of the ELK stack	
Characteristic	Count
Yes	23
No, total	169
-No, but only Elasticsearch addressed	131
-No, but only Logstash addressed	13
-No, but only Kibana addressed	1
-No, but all addressed	22
Total	214

Information source level of detail	
Characteristic	Count
Detailed	23
More detailed	36
Not detailed	150
Total	214

Related work: Limitations of related work

The literature synthesis has revealed the following limitations

- None of the analyzed scientific works highlighted explicitly the key features of the ELK stack adequately
- Present scientific works also neglect of investigating the applicability of the ELK stack for varying kinds of Big Data use cases
- None of the prevalent scientific works refers to the applicability of the ELK stack for Big Data use cases in the mobility industry

These enumerated limitations unveil a research gap and corroborate the need for this thesis, revealing the key features of the ELK stack on a holistic view and assessing its applicability for Big Data use cases in the mobility industry seems to be necessary.

Descriptive study: Related technologies

Related technologies in literature review

- Apache Solr is mentioned a few times as a related and competitor search engine to Elasticsearch
- Lucidworks provides an open source end-to-end solution, similar to the ELK stack, which is called Solr integrated with Logstash and Kibana (SiLK) stack
- Splunk Enterprise is an end-to-end solution and a commercial rival of Elasticsearch

Popular search engines

- Elasticsearch (1.), Apache Solr (2.), Splunk Enterprise (3.), MarkLogic (4.), and Sphinx (5.)

Qualitative descriptions and comparisons between Elasticsearch, Apache Solr and Splunk Enterprise, including SiLK stack.

Source: based on [7]

Descriptive study: Logstash

Logstash

- is a open-source tool engine
- provides an integrated framework for log collection, centralization, parsing, and analysis of a variety of structured and unstructured data
- is designed to efficiently and flexibly process logs, events, and unstructured data sources for distribution into a variety of outputs
- can be easily customized via plugins for input, output and data-filters
- is most commonly used to index data in Elasticsearch

Sources: based on [6, 8, 9, 10, 11, 12]

Descriptive study: Elasticsearch

Elasticsearch

- is a distributed and highly scalable open-source full-text search engine
- is a fairly new project that is built on top of Lucene
- goes beyond free-text search and provides structured search, hit word highlighting, aggregations, and facets over the data
- performs various types of searches and aggregations
- is primarily designed as a search engine
- has been given functionalities to act as a data storage solution
- is the main component in the ELK stack and provides its storage and search engine capabilities

Sources: based on [10, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23]

Descriptive study: Kibana

Kibana

- is the open-source front end system in the ELK stack
- is a data analysis portal that interacts with the Elasticsearch RESTful interface to retrieve data from Elasticsearch
- is based on HTML, JavaScript, and Bootstrap
- requires a web server, included in the Kibana 4 package, and it is fully compatible with any modern browser
- is not a requirement for querying the search cluster
- supports time-based comparisons, easy creation of graphical data representations like plots, charts and maps, flexible and responsive web interface, and a powerful search syntax
- does not provide any authentication or authorization mechanism by default

Sources: based on [9, 12, 19, 21, 24, 25, 26, 27, 28, 29]

Case study: Overview of Big Data use cases (I/III)

Use case ID		UC-1	UC-2	UC-3	UC-4	UC-5	UC-6
Current solutions	Compute (system)	Linux servers	Linux servers	Linux servers	Linux servers	Linux servers	Linux servers
	Storage	HDFS and Hive	Local storage	Local storage	HDFS and Hive	HDFS and Hive	HDFS and Hive
	Networking	10 gigabit Ethernet	10 gigabit Ethernet	10 gigabit Ethernet	10 gigabit Ethernet	10 gigabit Ethernet	10 gigabit Ethernet
	Software	Hadoop, Hive, Python, and Tableau	Excel, Python, R, and Tableau	Excel and Qlikview	Hadoop, Hive, Python, Scoop, and Tableau	Hadoop, Hive, Excel, and Tableau	Excel, Hadoop, Hbase, Hive, Python, QlikView, R, Spark, and Tableau
Big data characteristics	Data source	Commu-nication data from vehicles, i.e. Internet of Things	Web surveys	Surveys	Production systems and production line computers	Call center databases and ticketing systems	Diagnostic/ technical data of vehicles, i.e. Internet of Things
	Data volume	High	Low	Low	Medium	Low	High
	Data velocity	High	Low	Low	Medium	Medium	High
	Data variety	High	Low	Low	Medium	Medium	High
	Data variability	None	Low	Low	Low	Low	High

Case study: Overview of Big Data use cases (II/III)

Use case ID		UC-1	UC-2	UC-3	UC-4	UC-5	UC-6
Big data science	Data veracity	Unknown	Low	Low	Medium	Low	Medium
	Visualization	Python and Tableau	Excel and Tableau	Excel and QlikView	Tableau	Excel and Tableau	Excel, QlikView, and Tableau
	Data types	Machine-generated streaming car data, XML, and structured (automatic generated)	Textual, XLSX, and structured (manually generated)	Textual quality/warranty data, CSV, and structured	Production data, XML, and structured (automatic generated)	Textual, CSV, and structured and unstructured (semi-automatic generated)	Machine-generated streaming car data, XML, and unstructured (automatic generated)
	Data analytics	Descriptive analytics: yes Diagnostic analytics: no Predictive analytics: no Prescriptive analytics: no	Descriptive analytics: yes Diagnostic analytics: yes Predictive analytics: yes Prescriptive analytics: no	Descriptive analytics: yes Diagnostic analytics: no Predictive analytics: no Prescriptive analytics: no	Descriptive analytics: yes Diagnostic analytics: yes Predictive analytics: yes Prescriptive analytics: no	Descriptive analytics: yes Diagnostic analytics: yes Predictive analytics: yes Prescriptive analytics: no	Descriptive analytics: yes Diagnostic analytics: yes Predictive analytics: yes Prescriptive analytics: no

Case study: Overview of Big Data use cases (III/III)

Use case ID	UC-1	UC-2	UC-3	UC-4	UC-5	UC-6
Big data specific challenges (gaps)	Data volume and velocity requirements and up-to-date analytics	None	None	Data consolidation: Merging multiple data sources and cleaning of human entered and poorly encoded data	Unstructured textual data	Data is highly unstructured, dynamic, and has different hierarchical structures
Security and privacy requirements	Sensitive data, i.e. location-based data	Sensitive data, i.e. customer-specific data	Sensitive data, i.e. customer-specific data	Sensitive data, i.e. production-related data	Sensitive data, i.e. caller data	Sensitive data, i.e. vehicle-identification information
Highlight issues for generalizing this use case	Moving from batch analytics to real-time streaming analytics	None	None	Moving from batch analytics to real-time streaming analytics	Text mining requirements	Real-time data streaming issue Moving from batch analytics to real-time streaming analytics

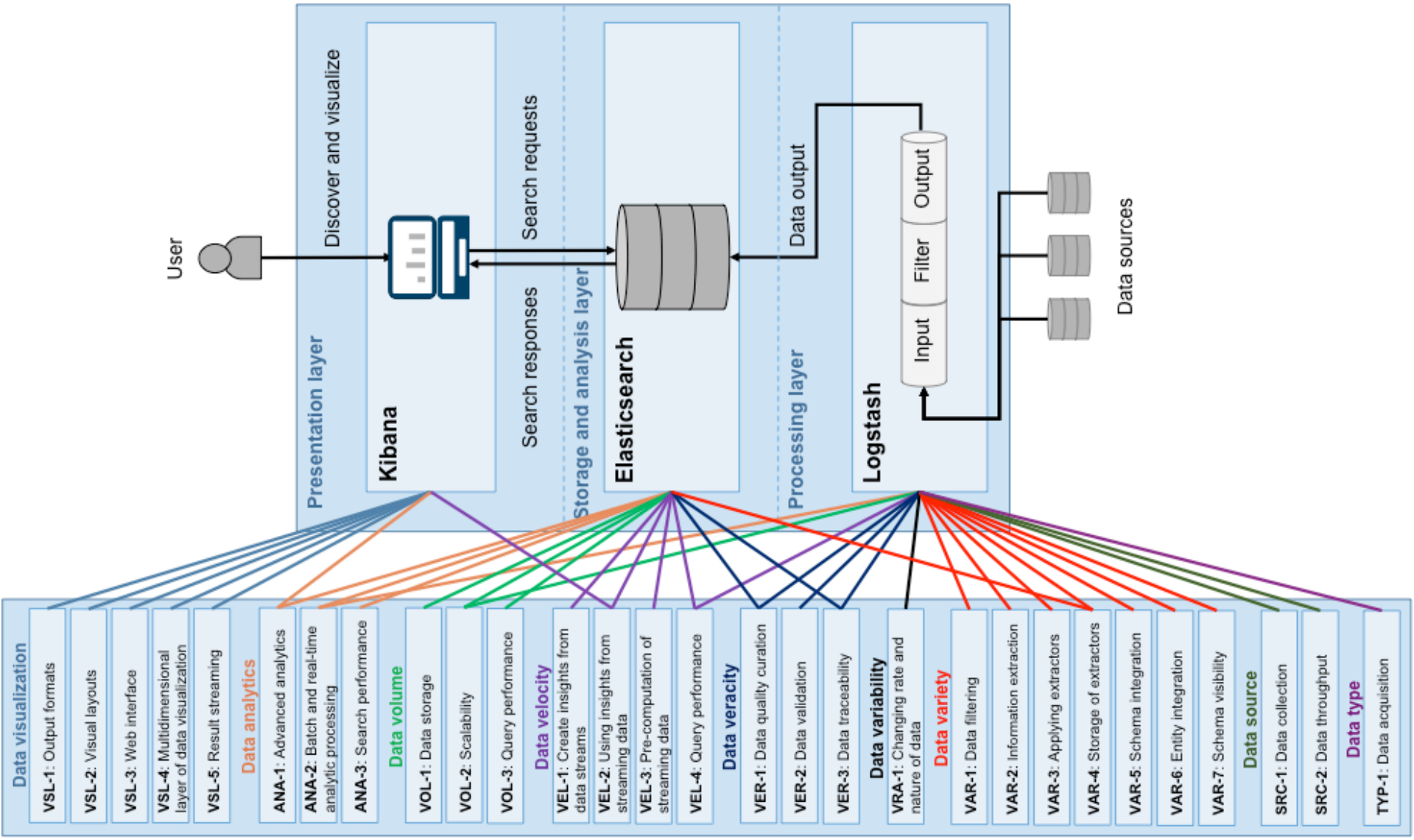
Cross-case analysis: Requirements and ELK stack mapping

sebis

Requirements categories:

- Data source (2)
- Data volume (3)
- Data velocity (4)
- Data variety (7)
- Data variability (8)
- Data veracity (3)
- Data visualization (5)
- Data type (1)
- Data analytics (3)

Cross-case analysis: Requirements and ELK stack mapping



Cross-case analysis: Example of a Big Data use case requirement description

Requirement overview	
Id	SRC-1
Name	Data collection
Type	Functional
Category	Data source
Description	The system shall support reliable real-time, asynchronous, streaming, and batch processing to collect data from various data sources.
Literature	[30]

Cross-case analysis: Excerpt of an assessment

UC-1			
Requirement category	**Requirement ID and requirement name**	**Instantiated requirement**	**Technology assessment**
Data source	SRC-1: Data collection	The vehicle communication data has to processed in reliable real-time from the data sources into the cluster.	Logstash: With the help of the Logstash Forwarder, the specific directory of the remote server is listened for new incoming communication data. With the arrival of new communication data, the Logstash Forwarder directly forwards this data to the Logstash central server. Within the central server, the communication data is parsed very fast and sent directly to Elasticsearch.
Data source	SRC-2: Data throughput	The system shall enable high-throughput data transmission between the data source and computing cluster.	Logstash: The Logstash Forwarder is able to forward several hundreds of events per second. Also during the implementation, the Logstash Forwarder was able to forward the data during a peak time of incoming events without queuing the incoming data. Logstash is able to parse the forwarded data without queuing into Elasticsearch. Based on the extent of the filter operations, the amount of forwarded data cannot be processed directly, since complex filter operations take more time and new forwarded data can be queued in the filter until the filter operations are finished.

Cross-case analysis: Assessment summary table

Category	Requirement	UC-1	UC-2	UC-3	UC-4	UC-5	UC-6	Problem
Data source	SRC-1: Real-time and batch processing	L			L	L	L	
	SRC-2: Data throughput	L			L	L	L	
Data volume	VOL-1: Data storage	E	E	E	E	E	E	
	VOL-2: Scalability	E			E	E	E	
	VOL-3: Query performance	E, L			E, L	E, L	E, L	
Data velocity	VEL-1: Create insights from data streams	E						
	VEL-2: Using insights from streaming data	E, K						Real-time
	VEL-3: Pre-computation of streaming data	E						
	VEL-4: Query performance	E, L						
Data variety	VAR-1: Data filtering	L			L		L	
	VAR-2: Information extraction	L	L	L	L	L	L	
	VAR-3: Applying extractors	L			L		L	
	VAR-4: Storage of extractors	E, L	E, L	E, L	E, L		E, L	
	VAR-5: Schema integration							
	VAR-6: Entity integration							
	VAR-7: Schema visibility							
Data variability	VRA-1: Changing rate and nature of data				L	L	L	Changing hierarchies in XML files
Data veracity	VER-1: Data quality curation	E, L	E, L	E, L	E, L	E, L	E, L	
	VER-2: Data validation							
	VER-3: Data traceability				E, L	E, L	E, L	
Data visualization	VSL-1: Output formats				K	K	K	
	VSL-2: Visual layouts				K	K	K	
	VSL-3: Web interface	K	K	K	K	K	K	
	VSL-4: Multidimensional layer of data visualization				K	K	K	
	VSL-5: Result streaming							
Data type	TYP-1: Data acquistion	L	L	L	L	L	L	
Data analytics	ANA-1: Advanced analytics	E, K	E, K	E, K	E, K	E, K	E, K	Type of data analytics capability
	ANA-2: Batch and real-time analytic processing	E, L			E, L	E, L	E, L	Real-time
	ANA-3: Search performance	E	E	E	E	E	E	

Cross-case analysis: Key findings

The ELK stack

- provides near real-time data analytics capabilities
- provides descriptive data analytics capabilities
- satisfies most of the requirements in regard to the 5Vs of Big Data
  - especially data volume and data veracity
- has difficulties with processing highly variable and multiple hierarchy XML files
  - Logstash's processing capabilities are limited for this kind of data

When to use the ELK stack?

- for understanding the data (see Cross Industry Standard Process for Data Mining, or CRISP-DM)
- in descriptive and explorative Big Data use cases

When not to use the ELK stack?

- in Big Data use cases which only require diagnostic, predictive, or prescriptive data analytics capabilities
- in Big Data use cases where data have to be processed and analyzed within milliseconds

Conclusion and outlook: Limitations

Limitations of the master's thesis

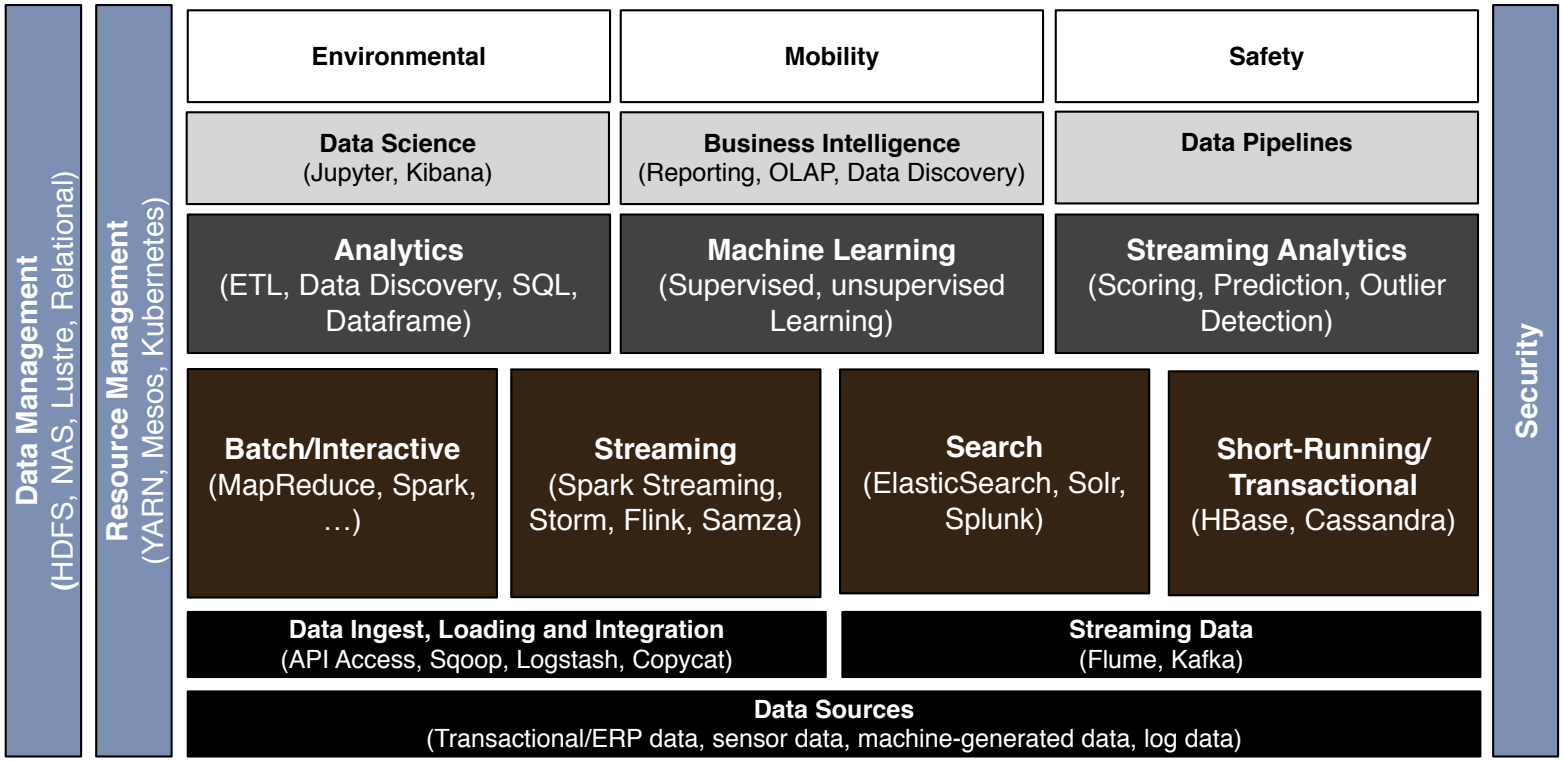
- Assessment of the ELK stack is based only a fractional amount of big data use cases in the mobility industry
- Case studies are only based on Big Data use cases of only one company
- Focus on qualitative evaluation of the ELK stack applicability
- Cross-case analysis neglects security requirements
- Isolated view on the ELK stack without analyzing its role within a Big Data workbench
- Missing assessment of the integration with other Big Data technologies, e.g., Apache Hadoop

Conclusion and outlook: Conclusion

- The ELK stack is able to collect, parse, and analyze various kinds of structured, semi-structured, and unstructured data from various input sources in near real-time
- It provides a rich web interface for descriptive data analytics
- The ELK stack encounters the requirements of Big Data use cases very well, but does not provide sufficient capabilities for real-time and advanced analytics
- The ELK stack is as strong as its weakest technology

Outlook for future research

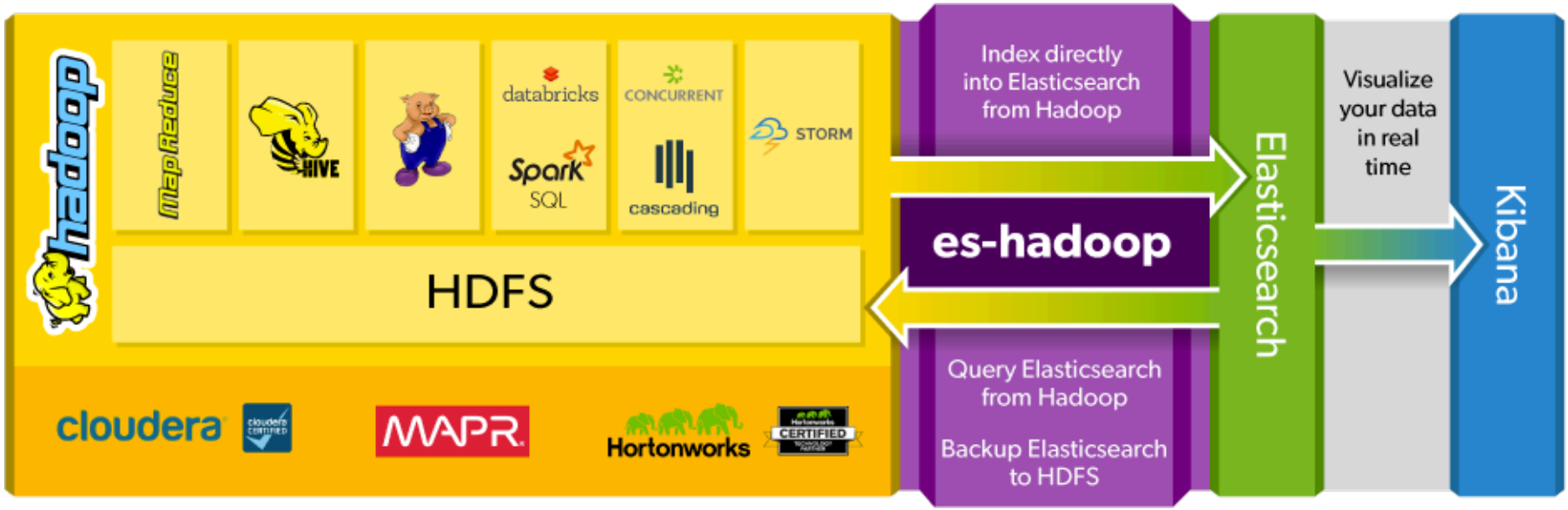
- Analyzing the role of the ELK stack within a Big Data workbench



Source: based on [31]

Outlook for future research

- Assessing the integration of the ELK stack with the Apache Hadoop ecosystem



Source: based on [32]

Near real-time Twitter analysis

Thank you very much for your attention!

Do you have any questions?

References

[1] Michael Friedewald and Oliver Raabe. 2010. Ubiquitous computing: An overview of technology impacts. Telematics Inform, 28, 2 (2011), 55 – 65.

[2] Günther Sagl, Martin Loidl, and Bernd Resch 2012. Visuelle Analyse von Mobilfunkdaten zur Charakterisierung Urbaner Mobilität. In *Geoinformationssysteme*, Wichmann Verlag, Berlin, Germany, 72 – 79.

[3] Andre Luckow, Ken Kennedy, Fabian Manhardt, Emil Djerekarov, Bennie Vorster and Amy Apon. 2015 In Proceedings of IEEE Conference on Big Data. IEEE. Santa Clara, CA, USA.

[4] An Introduction to the ELK stack. Elastic. http://www.elastic.co/webinars/introduction-elk-stack. Accessed: 2016-02-27.

[5] The ELK Stack in a DevOps Environment. Elastic. http://www.elastic.co/webinars/elk-stack-devops-environment. Accessed: 2016-02-27.

[6] Saurabh Chhajed. Learning ELK Stack. Packt Publishing, 2015.

[7] DB-Engines. Db-engines ranking of search engines. http://db-engines.com/en/ranking/search+engine. Accessed: 2016-02-27.

[8] S. Bagnasco , D. Berzano, A. Guarise, S. Lusso, M. Masera, and S. Vallero. Monitoring of iaas and scientific applications on the cloud using the elasticsearch ecosystem. In Journal of Physics: Conference Series, volume 608, page 012016. IOP Publishing, 2015.

[9] Zirije Hasani, Boro Jakimovski, Margita Kon-Popovska, and Goran Velinov. Real time analytic of sql queries based on log analytic. In ICT Innovations 2015 Web Proceedings, pages 78 – 87, Ohridnia, 2015.

[10] Kurt Hurtado and Tal Levy. Going beyond the needle in a haystack: Elasticsearch and the elk stack. https://www.speakerd.s3.amazonaws.com/presentations/70d4390b68504e02b4a40a1aa3754532/strata15-elk-stack-needle-haystack.pdf. Accessed: 2016-02-27.

[11] Chris Sanders and Jason Smith. Applied Network Security Monitoring. Syngress, 2013.

[12] James Turnbull. The Logstash Book Version 2.0. 2015.

[13] Vahid Abbasi. Phonetic analysis and searching with google glass api. Master's thesis, Uppsala Universitet, 2015.

[14] Radu Hambasan, Michael Kohlhase, and Corneliu Prodescu. Mathwebsearch at ntcir-11. In Proceedings of the 11th NTCIR Conference, Tokyo, Japan, 2014. NTCIR.

[15] Kerim Meijer, Manos Tsagkias, and Magiel Bruntink. Efficiently storing and searching distributed data. Master's thesis, University of Amsterdam, 2013.

References

[16] Manda Sai Divya and Shiv Kumar Goyal. Elasticsearch: An advanced and quick search technique to handle voluminous data. COMPUSOFT, An international journal of advanced computer technology, 2(6):171 – 175, 2013.

[17] Radoslav Bodoa and Daniel Kouril. Efficient management of system logs using a cloud. In The International Symposium on Grids and Clouds (ISGC) 2013, Taipei, Taiwan, 2013. Proceedings of Science.

[18] Jose Arias Fernandez, Quentin Bahers, Alberto Blazquez Rodrıguez, Marten Blomberg, Carl Carenvall, Kristian Ionescu, Sukhpreet Singh Kalra, Georgios Koutsoumpakis, Hao Li, Tommy Mattsson, Andreas Moregard Haubenwaller, Anders Steinrud, Tomas Savstrom, and Gabriel Tholsgard. Iot-framework. 2014.

[19] Jun Bai. Feasibility analysis of big log data real time search based on hbase and elasticsearch. In Natural Computation (ICNC), 2013 Ninth International Conference on, pages 1166 – 1170, Shenyang, China, 2013. IEEE.

[20] Johannes Stoll. Development of a distributed software architecture for the search and analysis of open data. Master's thesis, Hochschule Offenburg, 2014.

[21] Morten A. Iversen. When logs become big data. Master's thesis, University of Oslo, 2015.

[22] Juan Luis Perez. Desing and evaluation of scalable iot event processing platform. Master's thesis, Universitat Politecnica de Catalunya, 2014.

[23] Oleksii Kononenko, Olga Baysal, Reid Holmes, and Michael W. Godfrey. Mining modern repositories with elasticsearch. In Proceedings of the 11th Working Conference on Mining Software Repositories, pages 328 – 331. ACM, 2014.

[24] Daniel Cea, Jordi Nin, Ruben Tous, Jordi Torres, and Eduard Ayguade. Towards the cloudification of the social networks analytics. In Modeling Decisions for Artificial Intelligence, pages 192 – 203. Springer, 2014.

[25] Rob Appleyard and James Adams. Using the elk stack for castor application logging at ral. In The International Symposium on Grids and Clouds (ISGC) 2015, Taipei, Taiwan, 2015. Proceedings of Science.

[26] Mathias Knudsen, Jostein Laten, and Anders Dalmo. Deploying a virtualised high-interaction honeynet. Bachelor thesis, GjUniversity College, 2014.

[27] Timothy Riley. Detecting potentiallz compromised credentials in a large-scale production single-signon system. Master's thesis, Naval Postgraduate School, 2014.

[28] Marc Reilly. Is virtualisation the most secure way to provide shared resources and applications. Master's thesis, National College of Ireland, 2013.

[29] Yuvraj Gupta. Elasticsearch Server Second Edition. Packt Publishing, 2015.

[30] Wo Chang. Nist big data interoperability framework: Volume 6, reference architecturehttp://www.bigdatawg.nist.gov/_uploadfiles/NIST.SP.1500-6.pdf, 2015. Accessed: 2016-02-27.

References

[31] Mashrur Chowdhury, Amy Apon, Kakan Dey, Venkat Gudivada, Pradip Srimani, Beth Plale, Andre Luckow, Hongxin Hu, Chad Steed, John McGregor, Yuanchang Xie, Linh Ngo, Jason Hallstrom, and Jason Thatcher. Data Analytics in Connected Transportation Systems, 2016.

[32] Elasticsearch for Hadoop. Elastic. https://www.elastic.co/products/hadoop. Accessed: 2016-02-27.

[33] Philip Russom. 2011. Big Data Analytics. TDWI Best Practices Report, Fourth Quarter. The Data Warehouse Institute, Renton, WA, USA.

[34] Yuri Demchenko, Paola Grosso, Cees de Laat, and Peter Membrey. 2013. Addressing Big Data Issues in Scientific Data Infrastructure. In Proceedings of the 2013 International Conference on Collaboration Technologies and Systems. IEEE. San Diego, CA, USA, 48–55.

[35] Bart De Muynck. 2015. How to Derive Value From Big Data in Transportation. Gartner Research.

[36] Sebastian Wedeniwski. Mobilitätsrevolution in der Automobilindustrie: Letzte Ausfahrt digital! Springer Vieweg, 2015.

[37] Neil Chandler. Agenda overview for analytics, business intelligence and performance management, 2014. Gartner, 2014.

[38] DB-Engines. Vergleich der systemeigenschaften elasticsearch vs. solr vs. splunk. http://db-engines.com/de/system/Elasticsearch%3BSolr%3BSplunk. Accessed: 2016-02-27.

[39] Sachin Handiekar and Anshul Johri. Apache Solr for Indexing Data. Packt Publishing, 2015.

[40] Marc Seeger. Building blocks of a scalable web crawler. Master's thesis, Stuttgart Media University, 2010.

[41] Isabell Grube. Enhancing invenio digital library with an external relevance ranking engine. Bachelor thesis, Hochschule Karlsruhe a Technik und Wirtschaft, 2012.

[42] Xiaoming Gao, Vaibhav Nachankar, and Judy Qiu. Experimenting lucene index on hbase in an hpc environment. In Proceedings of the first annual workshop on High performance computing meets databases, pages 25 – 28. ACM, 2011.

[43] Jan Korenn. Fulltext search in the database and in texts of social networks. Master's thesis, National College of Ireland, 2013.

[44] Markus Klose and Daniel Wrigley. Einführung in Apache Solr. O'Reilly Verlag, 2014.

[45] Alfredo Serafini. Apache Solr Beginner's Guide. Packt Publishing, 2013.

References

[46] Lucidworks. Lucidworks silk analytics. https://lucidworks.com/products/silk/. Accessed: 2016-02-27.

[47] Lucidworks. Silk. https://doc.lucidworks.com/assets/attachments/SiLK.pdf. Accessed: 2016-02-27.

[48] Alexander Miller and Dominik Lekar. Evaluation of analysis and visualization tools for performance data, 2014.

[49] James Miller. Mastering Splunk. Packt Publishing, 2014.

[50] Betsy Page Sigman. Splunk Essentials. Packt Publishing, 2015.

[51] Splunk. Splunk enterprise overview. http://docs.splunk.com/index.php?title=Documentation:Splunk:Overview:Whatsinthismanual:6.1beta&action=pdfbook. Accessed: 2016-02-27.

[52] Elastic. Filter plugins. https://www.elastic.co/guide/en/logstash/current/filter-plugins.html. Accessed: 2016-02-27.

[53] Elastic. Input plugins. https://www.elastic.co/guide/en/logstash/current/input-plugins.html. Accessed: 2016-02-27.

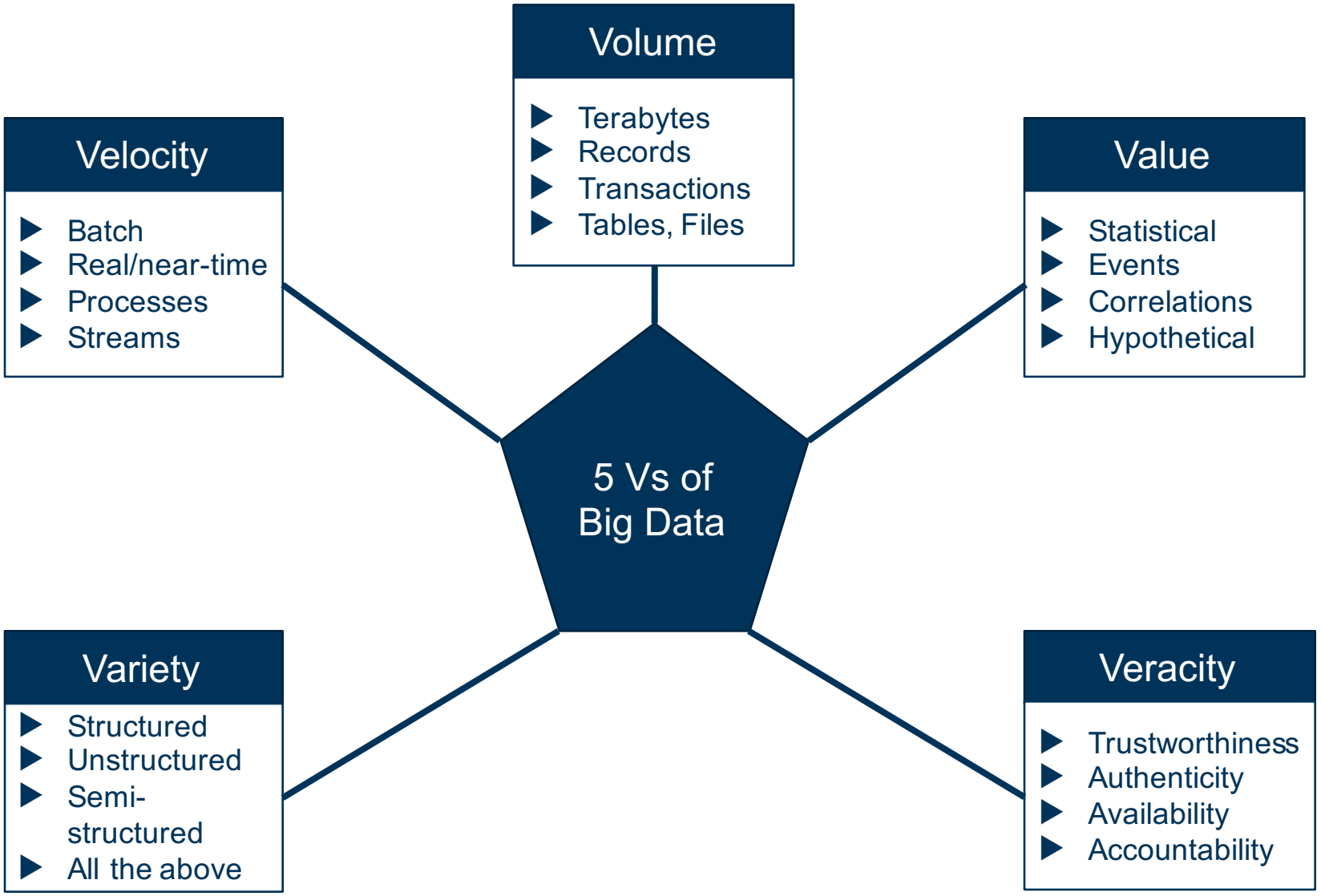
[54] Elastic. Output plugins. https://www.elastic.co/guide/en/logstash/current/output-plugins.html. Accessed: 2016-02-27.

[55] Francois Terrier. On elasticsearch performance. https://blog.liip.ch/archive/2013/07/19/on-elasticsearch-performance.html. Accessed: 2016-02-27.

[56] DB-Engines. Method of calculating the scores of the db-engines ranking. http://db-engines.com/en/ranking_definition. Accessed: 2016-02-27.

[57] Bernhard Pflugfelder. Real-time data analytics mit elasticsearch. https://www.inovex.de/fileadmin/files/Vortraege/real-time-data-analytics-mit-elasticsearch-bernhard-pflugfelder-jax2014.pdf. Accessed: 2016-02-27.

Appendix: Search-based data discovery tools



Sources: based on [33, 34]

Search-based data discovery tools

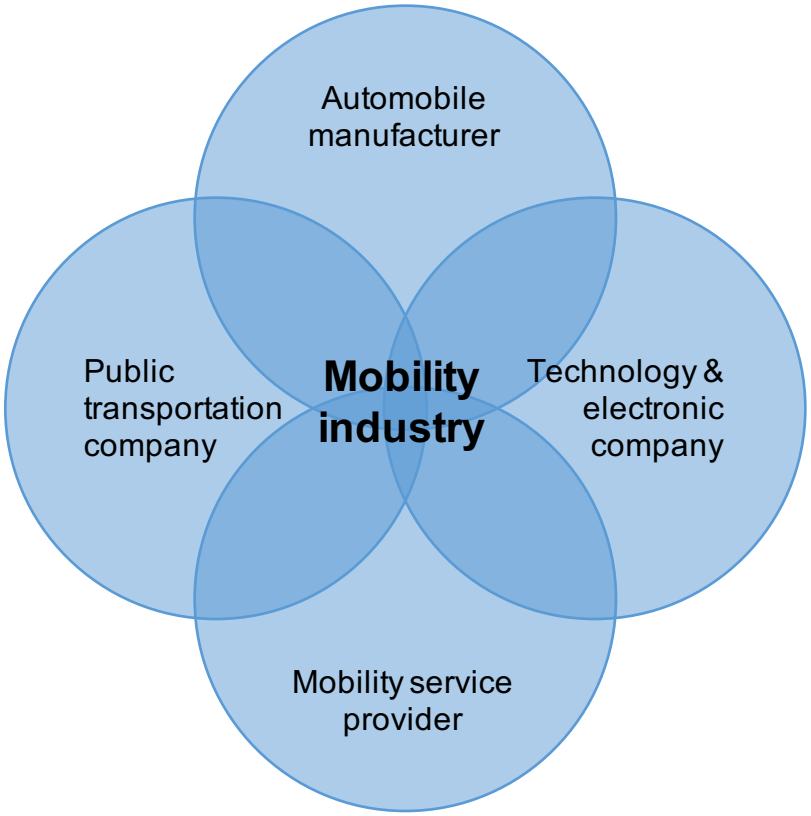
- raise huge expectations and promise high benefits for organizations among Big Data and analytics technologies.
- facilitate users to develop and refine views and analyses of multi-structured data using search term and to find relationships across structured, unstructured, and semi-structured data.
- feature a performance layer to lessen the need for aggregates and pre-calculations.
- are vended by i.e. Attivio, IBM, Oracle, Splunk, and ThoughtSpot

The combination of the three open source projects Elasticsearch, Logstash and Kibana (ELK), also known as the ELK stack is an outstanding alternative to commercial search-based data discovery tools.

Source: based on [35]

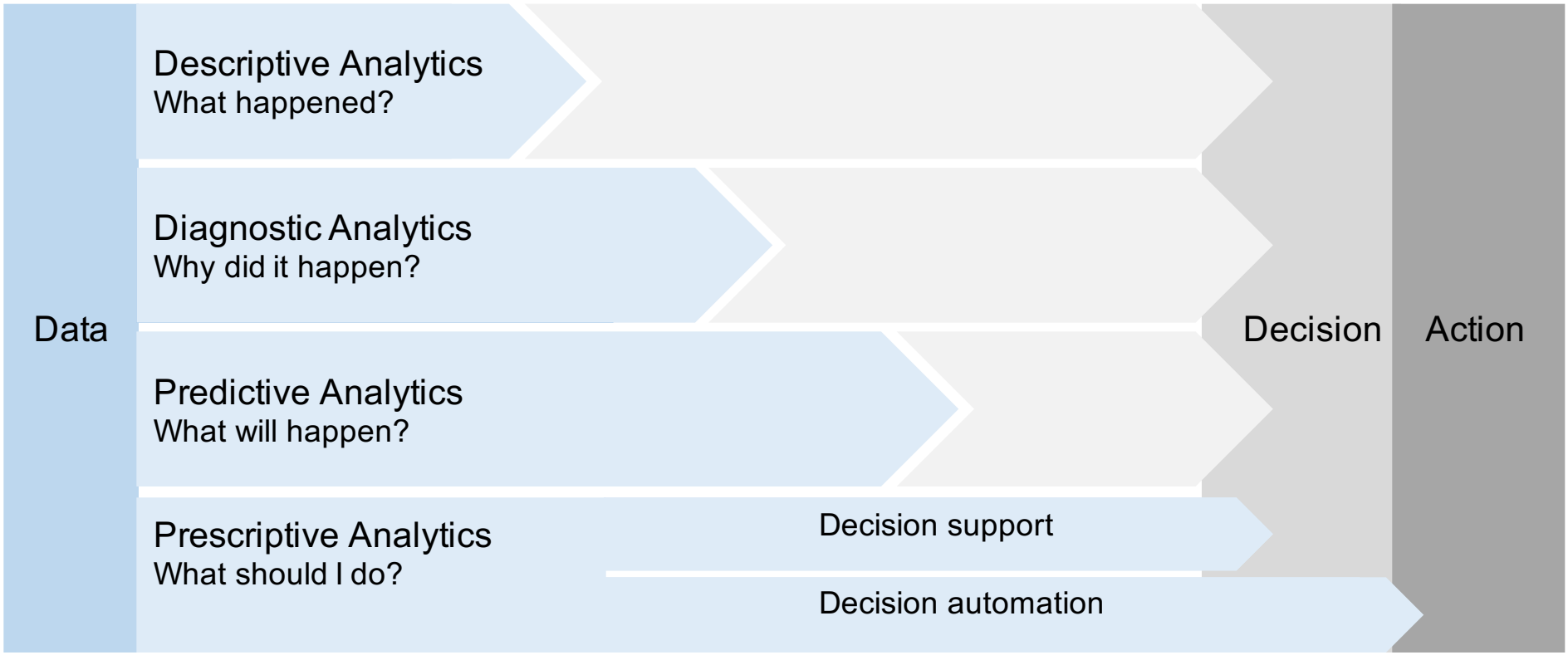
Appendix: Mobility industry

A company in the mobility industry makes a platform of integrated products and services available to overcome spatial distances dependent on individual needs. Essentially, the individual satisfaction of demand is optimized along the existing infrastructure for mobility. The mobility industry aims to adapt the underlying infrastructure according to the demand.



Source: based on [36]

Appendix: Four types of data analytics capability



Source: based on [37]

Appendix: Ranking of popular search engines

Ranking						
Rank			DBMS	Score		
Jan 16	Dec 15	Jan 15		Jan 16	Dec 15	Jan 15
1.	2.	2.	Elasticsearch	77.21	+0.65	+28.17
2.	1.	1.	Apache Solr	75.39	-3.75	-1.35
3.	3.	3.	Splunk Enterprise	43.12	-0.74	+10.05
4.	4.	5.	MarkLogic	9.92	-0.44	+0.89
5.	5.	4.	Sphinx	8.98	-0.02	-1.16

The calculation of the popularity ranking is based on specific parameters like the number of mentions of the system on websites, the general interest in the system, or the relevance in social networks (for the complete explanation of the popularity calculation, see [56])

Source: based on [7]

Appendix: Ranking of popular search engines

System properties comparison			
Name	**Elasticsearch**	**Solr**	**Splunk**
License	open source	open source	commercial
Implementation language	Java	Java	
Server operating systems	All OS with a JVM	All OS with a JVM and a servlet container	Linux, OS X, Solaris and Windows
Data scheme	schema-free	yes	yes
Secondary indices	yes	yes	yes
APIs and other access methods	Java API and RESTful HTTP/JSON API	Java API and RESTful HTTP API	HTTP REST
Supported programming languages	NET, Erlang, Java, JavaScript, Perl, PHP, Python, Ruby and Scala	.NET, Erlang, Java, JavaScript, XML, JSON, Perl, PHP, Python, Ruby and Scala	C#, Java, JavaScript, PHP, Python and Ruby
Partitioning methods	Sharding	Sharding	Sharding
MapReduce	no	no	
Consistency concepts	Eventual consistency	Eventual consistency	Eventual consistency
Transaction concepts	no	optimistic locking	no
Concurrency		yes	yes
User concepts			Access rights for users and roles

Source: based on [38]

Apache Solr

- is one of the most popular open source search platforms from the Apache Lucene open source project
- is written in Java and is built on top of Lucene, which offers core functionality for data indexing and search
- was initially started in 2004 at CNET
- uses a NoSQL-like document store database system
- extends Lucene by providing many useful features related to full-text search, e.g., keyword highlighting, spelling suggestions, complex ranking options, geospatial search or numeric field statistics
- also includes near real-time indexing, dynamic clustering, query language extension, caching and rich document handling
- supports distributed indexing by its SolrCloud technique

Sources: based on [39, 40, 41, 42, 43]

Appendix: Apache Solr

Key features

- Advanced full-text search
- Faceted search
- Spelling suggestions
- Language analysis
- Highlighting
- Near real-time search
- Multiple client APIs
- Scalability

Sources: based on [44, 45]

SiLK stack

- includes a custom packaging of Solr, Banana and a Solr Writer for Logstash
- is an analytics tool to analyze and visualize log data
- can be used for different use cases, such as for Apache weblogs or data analytics
- **Banana**
  - is the name of the open source port of Kibana 3
  - is a data visualization tool that allows to create dashboards to display content stored in Solr indices
  - provides panels such as histograms, geomaps, heatmaps and bettermaps for analyzing data
- **Solr**
  - stores data processed by Logstash
- **Solr Writer for Logstash**
  - is an implementation of Logstash specifically designed for indexing logs or other contents to Solr

Sources: based on [46, 47]

Splunk Enterprise

- is a log-, monitoring- and reporting tool
- manages searches, inserts, filters, and deletes, and analyzes Big Data that is created by machines, as well as other types of data
- has a free version that allows users to index up to 500 MB of data per day
- utilizes a role-based security model to offer flexible and effective ways to protect all the data indexed by Splunk, by controlling the searches and results in the presentation layer

Sources: based on [48, 49, 50]

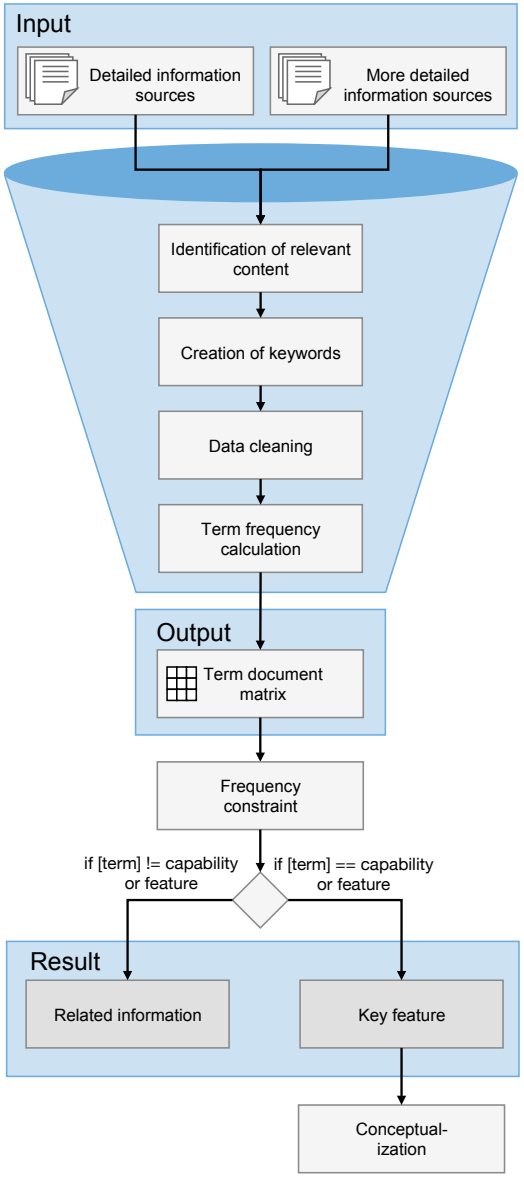
Appendix: Splunk Enterprise

Key features

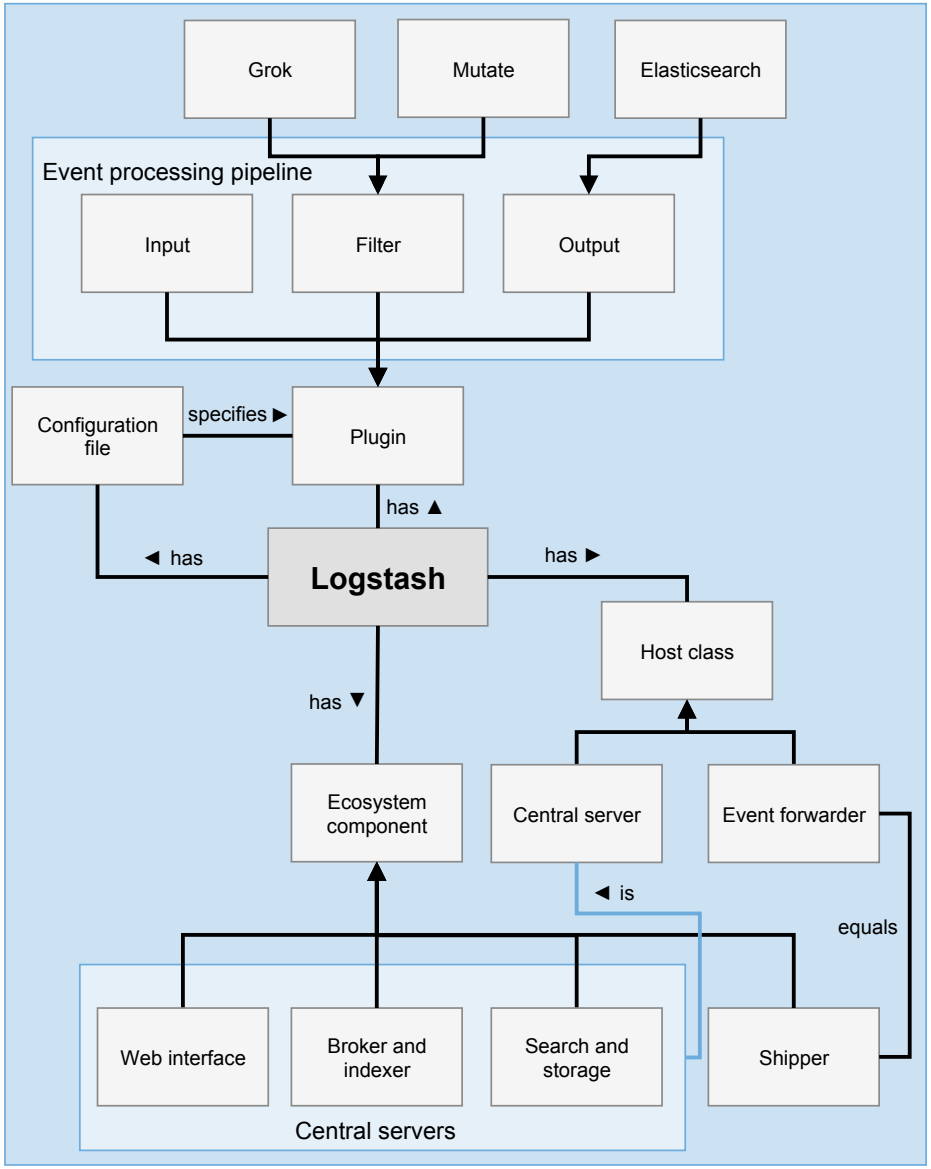
- Search
- Reports
- Dashboards
- Alerts

Source: based on [51]

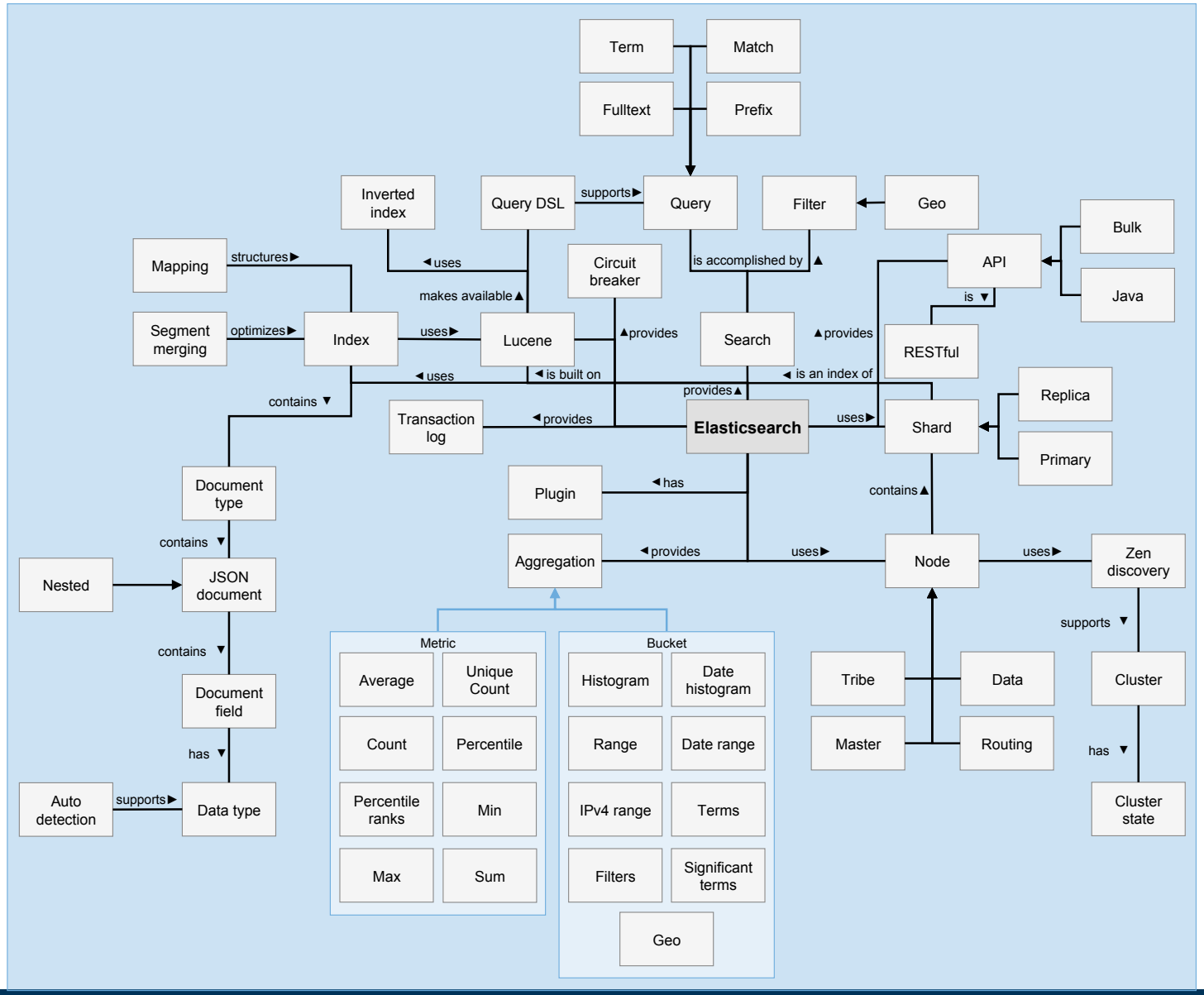
Appendix: Key feature extraction methodology



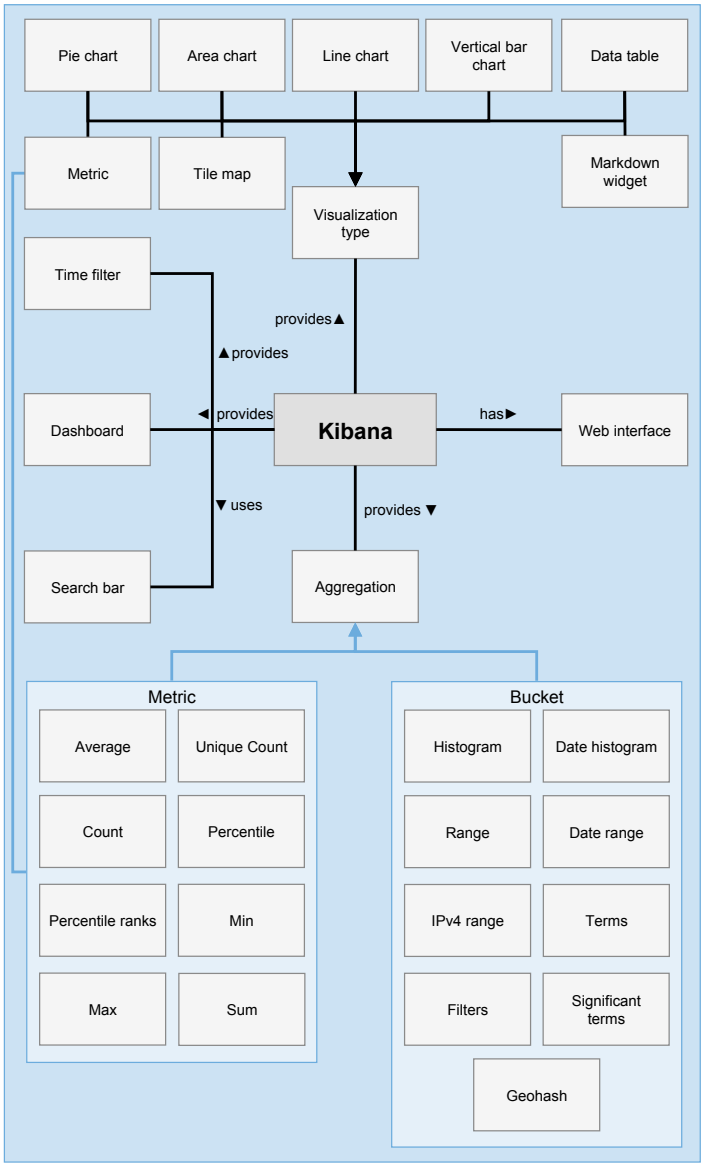
Appendix: Logstash conceptualization



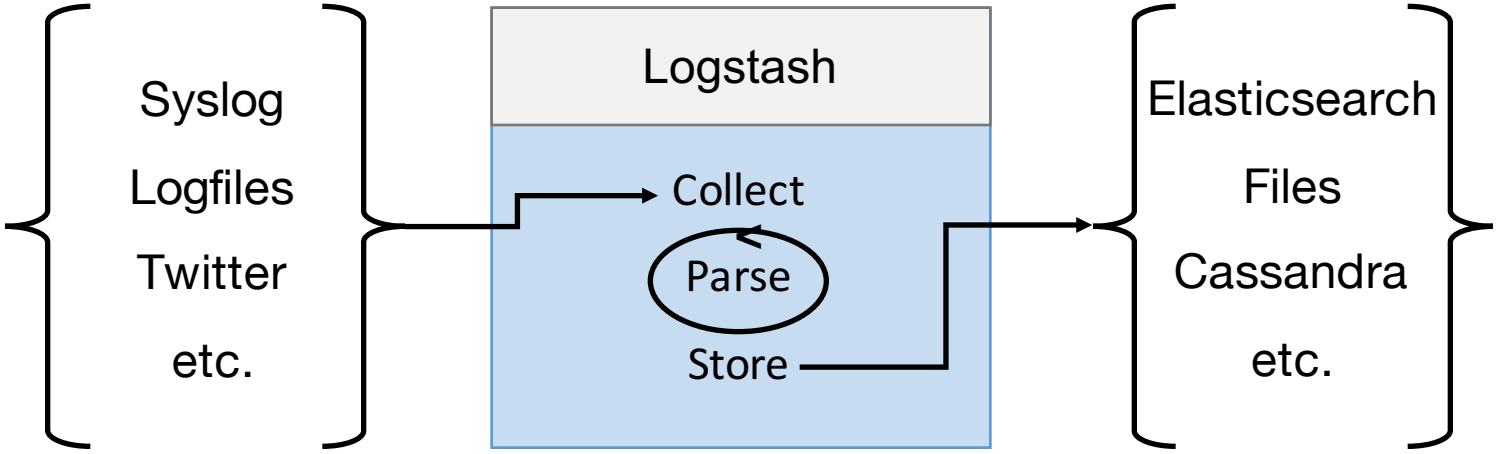
Appendix: Elasticsearch conceptualization



Appendix: Kibana conceptualization



Appendix: Logstash‘s event processing pipeline



Source: based on [21]

Appendix: Logstash plugins

Input	Filter	Output
file	csv	csv
jdbc	date	elasticsearch
lumberjack	geoip	email
kafka	grok	file
stdin	mutate	kafka
twitter	xml	stdout

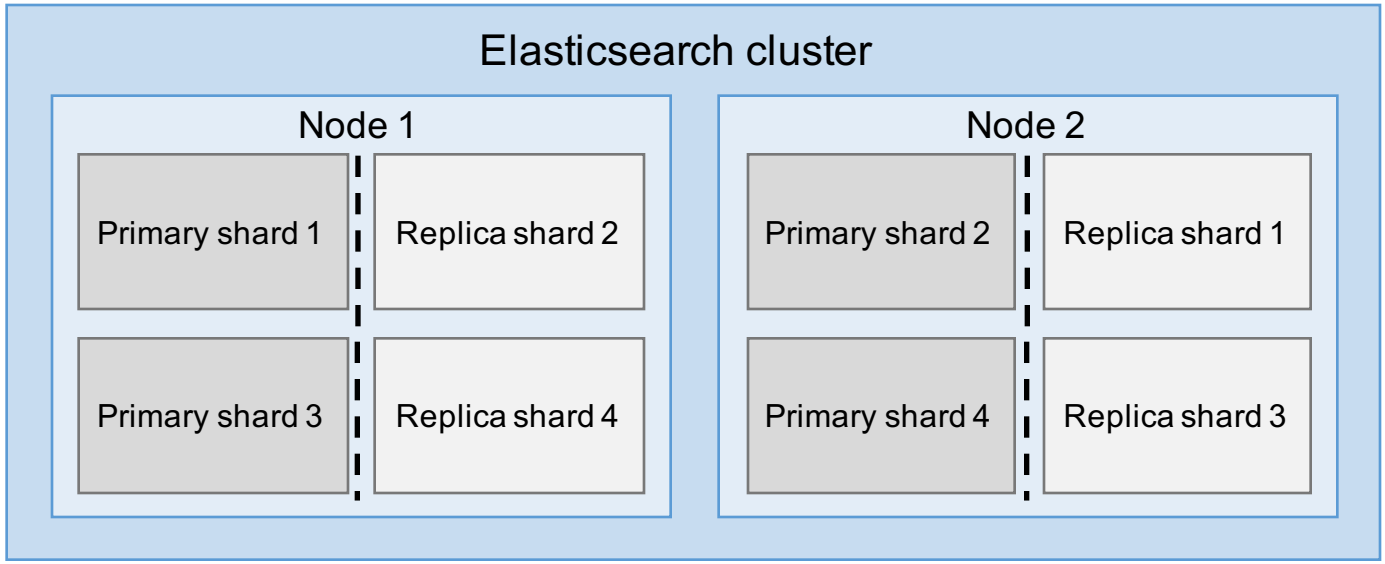
Sources: based on [52, 53, 54]

Appendix: Logstash's event output

logstash-2.1.1 — java — 82×46

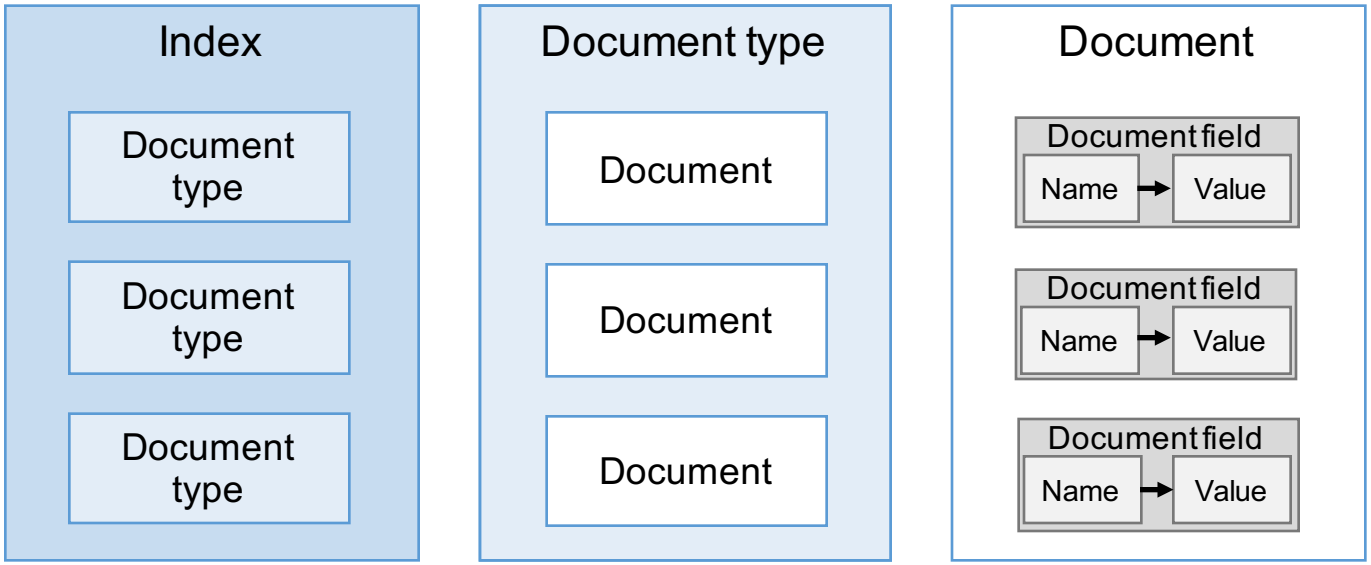

```
{
        "message" => "1,2000008,canSpeed,426987,0,9642,5710,M,26,MI_State_DOT,1FT
FX1EF2EKD69821,\"2014-09-10,15:15:00\",42.33691,-82.99779,471,8.986,m/s,20.1,mph,0
,,,,,,,,,,,,\r",
       "@version" => "1",
     "@timestamp" => "2014-09-10T19:15:00.000Z",
           "host" => "Oemers-MacBook-Pro-2.local",
           "path" => "/Users/oemeruludag/Desktop/rwd.csv",
       "SourceId" => "1",
     "ObsTypeID" => "2000008",
   "ObsTypeName" => "canSpeed",
      "SensorID" => "426987",
   "SensorIndex" => "0",
    "PlatformID" => "9642",
        "SiteID" => "5710",
      "Category" => "M",
     "ContribID" => "26",
   "Contributor" => "MI_State_DOT",
  "PlatformCode" => "1FTFX1EF2EKD69821",
     "Timestamp" => "2014-09-10,15:15:00",
      "Latitude" => 42.34,
     "Longitude" => -83.0,
     "Elevation" => "471",
   "Observation" => "8.986",
         "Units" => "m/s",
  "EnglishValue" => "20.1",
  "EnglishUnits" => "mph",
     "ConfValue" => "0",
        "Flag 1" => nil,
        "Flag 2" => nil,
        "Flag 3" => nil,
        "Flag 4" => nil,
        "Flag 5" => nil,
        "Flag 6" => nil,
        "Flag 7" => nil,
        "Flag 8" => nil,
        "Flag 9" => nil,
       "Flag 10" => nil,
       "Flag 11" => nil,
       "Flag 12" => nil,
       "Flag 13" => nil,
      "Location" => [
        [0] -83.0,
        [1] 42.34
    ]
}
```

Appendix: Elasticsearch cluster example



Source: based on [55]

Appendix: Elasticsearch data model



Source: based on [57]

Appendix: Elasticsearch vs. traditional relational database systems

sebis

Elasticsearch	Traditional relational database systems
Index	Database
Mapping	Schema
Document type	Table
JSON document	Row
Document field	Column

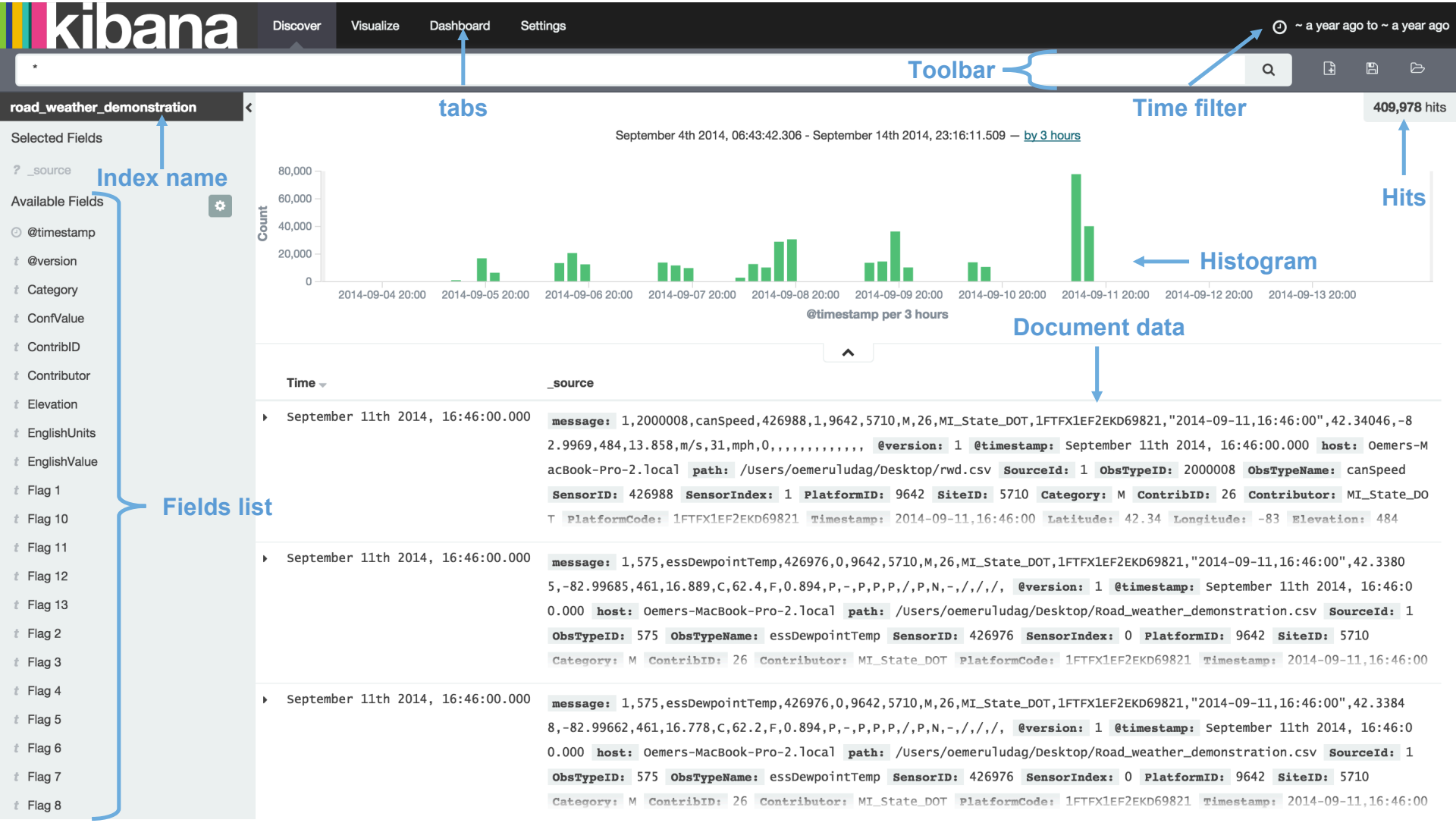
Appendix: Lucene‘s inverted index

1. The hills have eyes
2. Dawn of the dead
3. Shaun of the dead
4. Night of the living dead

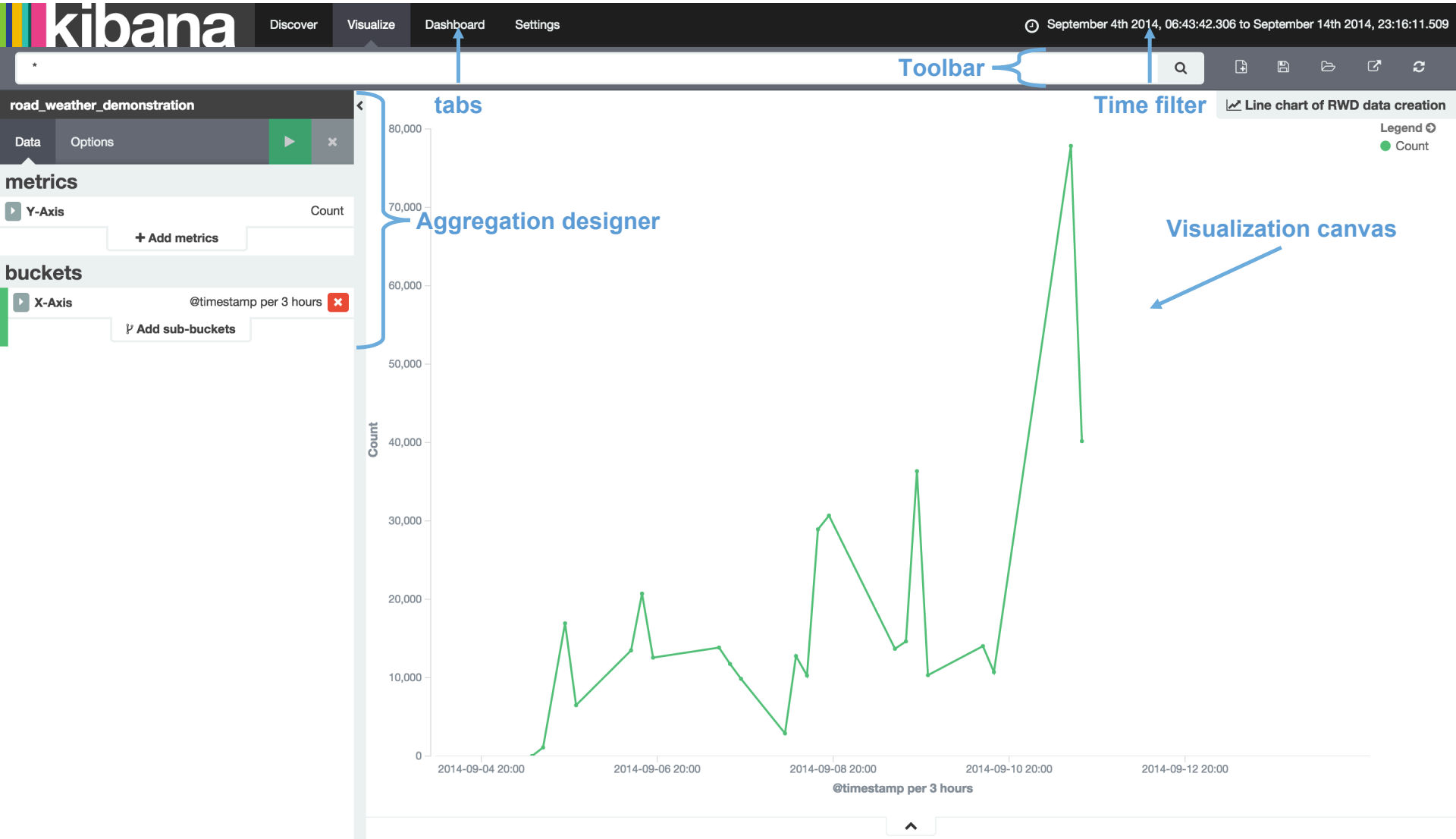
term	frequency	documents
• dawn	1	2
• dead	3	2,3,4
• eyes	1	1
• have	1	1
• hills	1	1
• living	1	4
• night	1	4
• of	3	2,3,4
• shaun	1	3
• the	4	1,2,3,4

Source: based on [21]

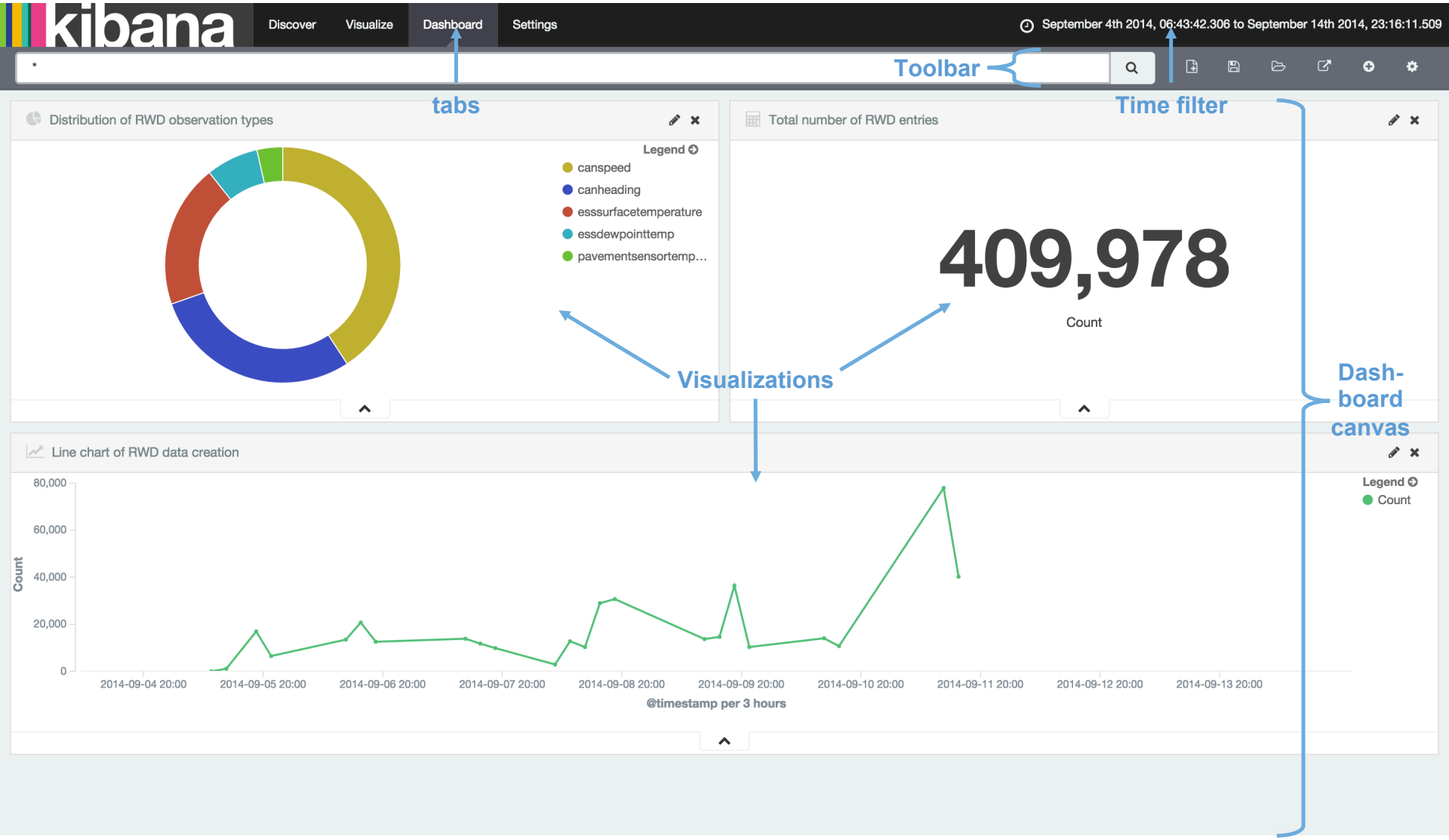
Appendix: Kibana's Discover page



Appendix: Kibana's Visualize page



Appendix: Kibana's Dashboard page



Appendix: Requirements elicitation methodology