Primary Key Associates Limited

Data Analytics

Data Analytics is at the core of Primary Key Associates work

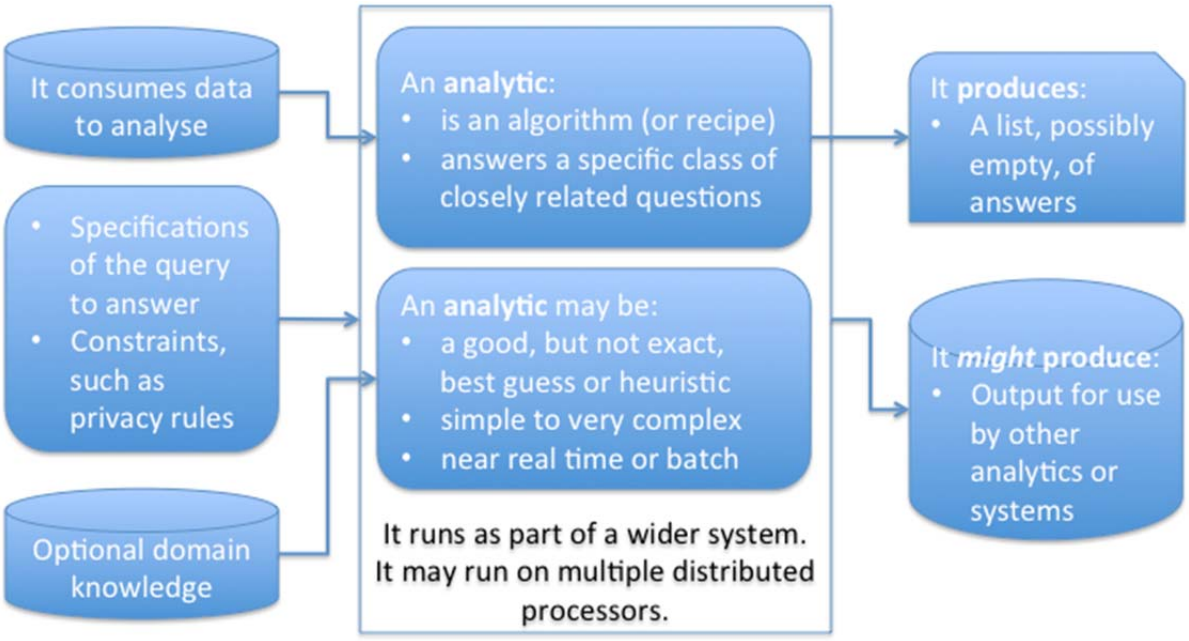
Our approach to analytics

In this paper Andrew Lea, our Technical Director in charge of Data Analytics, describes some of the paradigms, models, and techniques we have developed to allow us to accomplish remarkable things for our clients.

Defining an Analytic

Analytics analyse data in response to queries, subject to constraints, and produce lists of answers

We consider an analytic to have these components (some of which are optional):



Analytics consume and analyse data in response to queries, and produce answers. Although analytics do not of themselves own data, nor have a user interface, and are not a data visualisation, they are normally part of a system that does.

Primary Key analytics are frequently based on Artificial Intelligence techniques, and may operate on unstructured data, such as summarising text, as well as numbers.

Paradigms and Principles

Five business focused analytic paradigms

Based on years of experience we have developed five paradigms which guide our use of analytics, and in particular to keep them focused on solving the particular business problem at hand. These paradigms are:

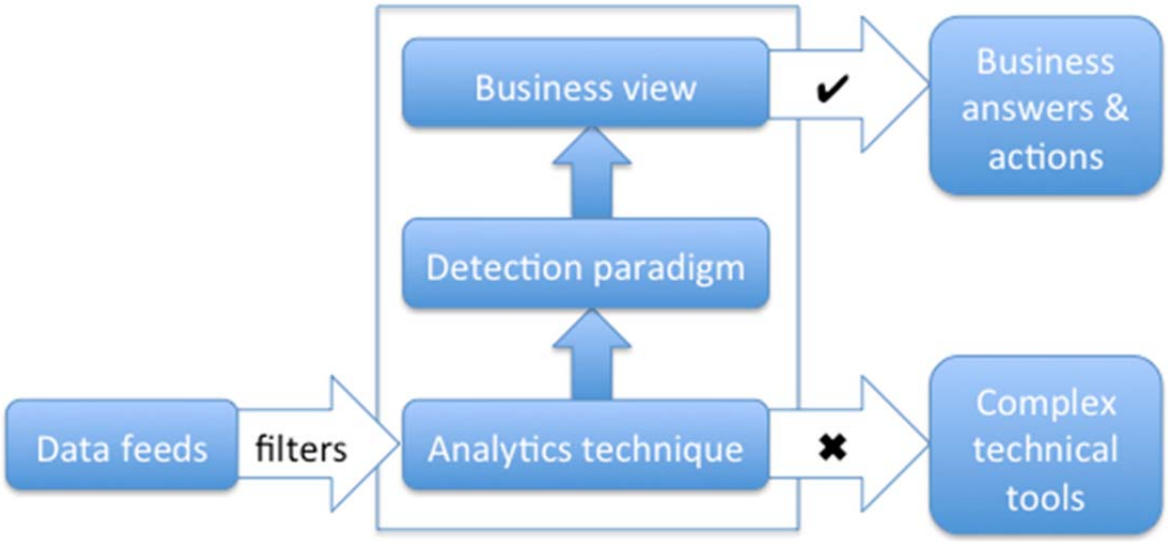
- Answer Based Analytics
- Scenario Based Analytics
- The Primary Key Principle: Minimal Complexity
- Minimum Infrastructure Impact
- Evidential Quality

Answer Based Analytics

Our analytics produce directly actionable intelligence

Many tools present an attractive visualisation of analysed data – such as social media - but still beg the question, what does this *mean* to the business? This question then has to be answered by analysts before feeding up to business leaders.

Our systems produce results that are immediately actionable, because they directly address business questions, supported with evidence. To know which questions to answer we work with clients to identify their interests, and the analytics which produce the answers they need. (If we have no such analytic, we develop one in short order.)



By way of example, Primary Key Distil is designed to turn data – consumer transactions, account data, transaction history – into actionable business answers, rather than the more typical complex technical reports, which then need to be further refined by analysts. [For more information, please see our paper on Fraud Discovery with Primary Key Distil].

Scenario Based Analytics

Describe the desired outcome, not the steps to find it

This recent refinement of Answer Based Analytics observes that people often cannot describe the criteria by which that which they are looking for can be located, but *would* be able to recognise it when they see it.

Using scenario based analytics we work with clients to develop 'for-instance' pictures of scenarios they would be concerned with, and use those pictures to develop scenarios which our analytics look for. We use an interesting declarative analytic language that we developed, in which the desired answer is modelled, and the system itself works out the steps to be taken to find it.

The Primary Key Principle: Minimal Complexity

Options produce complexity

Many systems give end-users multiple options to set which has several disadvantages:

- Different users with different settings, get different results, causing confusion.
- Users want to get their job done, rather than be experts in complex software.
- The end-user is often the person *least* qualified to know what settings should apply - especially when the developer could not decide what they should be.

We keep it simple, so it just works

The Primary Key Principle is therefore to:

- Have no settings for the user to adjust, and;
- Where settings are needed (under the hood, as it were), to decide them in software, handling the complexity ourselves.

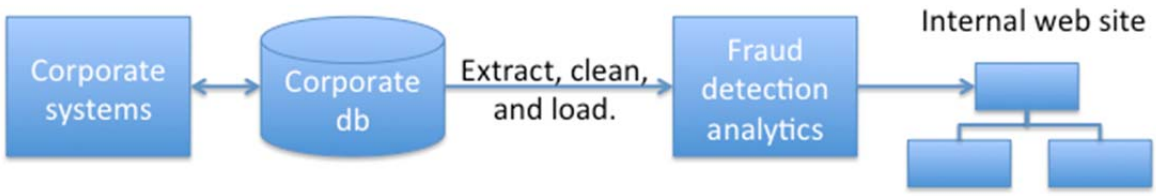
Minimum Infrastructure Impact

Lowest possible impact on your Information Systems

We design our systems and services to have minimum impact on your IT. If we are analysing your data, we always:

- load it read-only, so there is no possibility whatsoever of our software corrupting your existing data or interrupting your systems.
- Operate within your security perimeter, to ensure there can be no data loss.

For example, our fraud discovery process is generally deployed like this:



If we are providing data to you, we provide it via secure web sites, so no IT modifications are required. We know what IT departments prefer.

Evidential Quality Processing

High value results

Based on our many years' experience in capturing and analysing digital data for use in legal matters, we designed our systems to ensure that we carry out data acquisition and analysis to an evidential quality, so that it can be produced if necessary as part of legal proceedings.

Data and Data Acquisition

Different data presents different challenges

Data analytics clearly requires data to analyse, and indeed, it is appropriate when contemplating data analytics, to ask what data would best answer this problem? And what data do I have? Before data can be analysed, it must be imported. Different types of data present different challenges, but we have experience in analysing:

- Structured data – including transactions.
- Social media – for which we have our own advanced acquisition engines. [For more information on this, please see our Social Media Personnel Protection paper.]
- Natural language data – we can use not only metadata about text, but drawing on thirty years of natural language analysis can and normally do use the meaning within that text. In most languages.
- Image data – we can analyse both image metadata, for example for litigation, and content, in which we have experience in regard to space-craft remote sensing.
- Video data – again, we can analyse and improve video, for example, extracting good quality images from low quality video.

Data Storage

We store data safely, evidentially, and flexibly

Our data storage is underpinned by three principles:

1. We use our own storage infrastructure (not in 'the cloud') so that we can be sure where that data is for the whole of its lifecycle.
2. We store data that we acquire evidentially, so we can prove where it came from, and when.
3. We store data using a model that reflects the acquisition. Only when we come to analyse that data do we re-model it to the model on which that analytic is based. This helps keep us future proof.

Data Models

Data analytics models the world to make discoveries

To understand the world we live in, people make models of it. For example, the model we all carry in our heads of 'where a moving ball tends to go next' enables us to catch a ball. In a similar way, data analytics needs to have a 'model of the world' into which the data fits. This can either be explicit or, for simpler scenarios, implicit and unstated.

We have developed several models in which to analyse the world, which include:

- The Primary Key Seven Layer Model of Social Analytics
- The Three Entity Model
- The Bayesian Fraud Model (not discussed here)

The Primary Key Seven Layer Model of Social Analytics

Structure from chaos via seven layers

We analysis social media in terms of seven layers of actors, publications, and data:

1. **Population** – focuses on the overall ebb and flow of ideas
2. **Groups** – groups of actors with similar interests
3. **Communities** – actors who share interests, and are associated in communities
4. **Actors** – individual people, companies, web sites
5. **Publications** – individual publications, such as tweets or web pages
6. **Cleaned data** – identified publications, rendered in a common way
7. **Raw data** – the raw data acquired by the acquisition engines

The Three Entity Model

The human social world consists of people, events, and assets (which might be a place)

We designed our three entity model for conducting scenario based analytics on data which describes the interaction of people, web sites, companies, and so forth: in other words, the human social world.

We often prefer to use triples to represent data when using scenario-based analytics. (Triples are a particularly flexible way of representing data, and consists of subject-link-predicate. For example, Andrew Lea - works-at - Primary Key.) This model has three 'supra-entities', and mitigates a common trap of triples, namely the need to 'reify' or redesign triples when designs change.

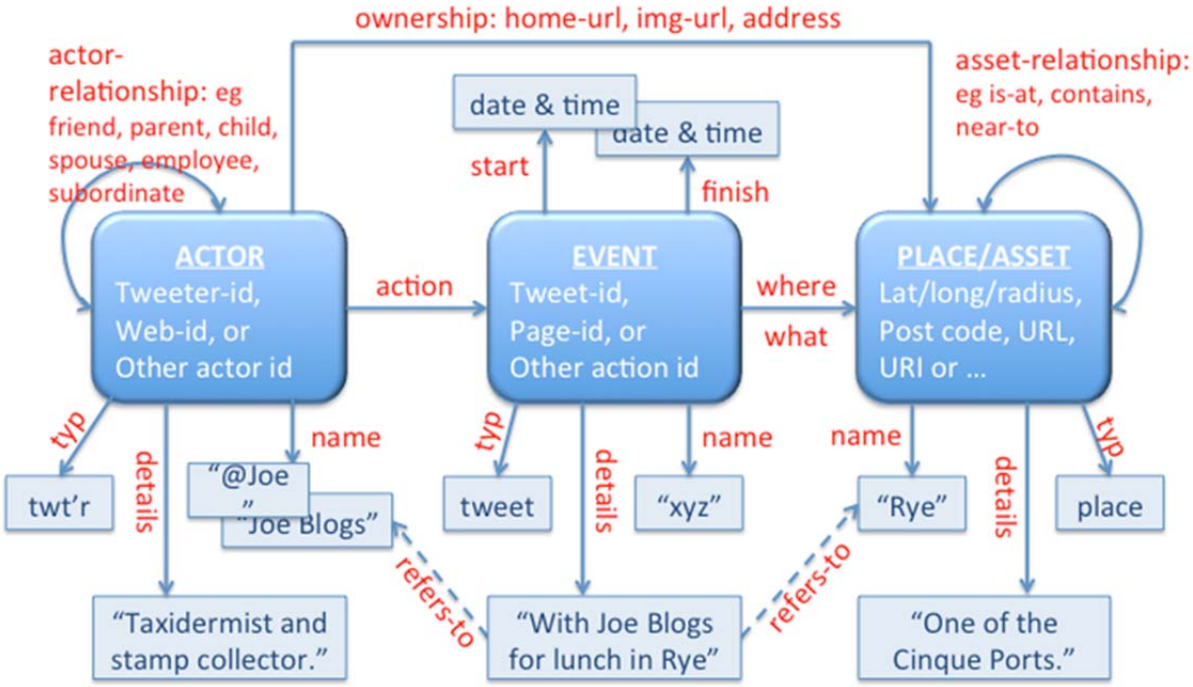
The three supra-entities are:

- **Actors**: people, web sites, companies – anyone who can do or say something. Actors participate in…
- **Events**: such as meetings, conversations, emails, tweets, which occur at a point in time either at or with:
- **Assets**: are things which can be owned, such as locations (or cyber-locations) or things, such as a car.

Each entity has very few properties, namely:

- **Id** – a unique reference
- **Typ** – what sort of entity it is
- **Name** – its name
- **Details** – one set of details, description, or text. Anything else is (in this model) better represented as a link to another entity.

Entities have links between each other, including within the same type. The three entity model looks like this:



Although it may look complex, it has reduced the world to only three classes of entity, four properties, and a few link types. We have not found a scenario we wish to model with it, but cannot.

Analysing Big Data

Smart and fast parallel processing

We overcome the challenge of 'big data' with three complementary approaches:

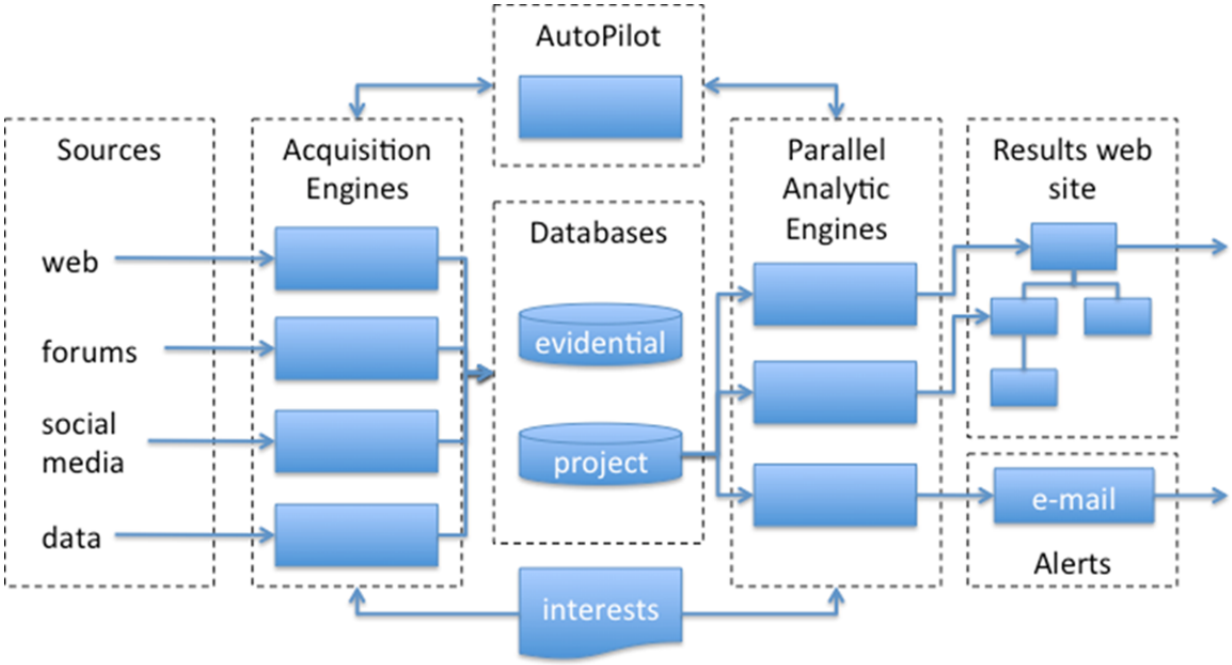
1. Parallel processing
2. Powerful integrated architectures
3. Smart algorithms – Order(N) processing

We use these approaches on a daily basis. Our Illuminate service, for example, uses them to enable complex questions about social media to be answered in a timely fashion. Without these approaches, our daily processing requirements would, we estimate, require about three months.

Parallel Processing

Parallel processing in Primary Key Illuminate

In general, we design our analytics to proceed in parallel. Our Primary Key Illuminate service uses this parallel architecture, which is supported by the highly virtualised infrastructure that we run, in which CPU cores can be re-allocated as required.

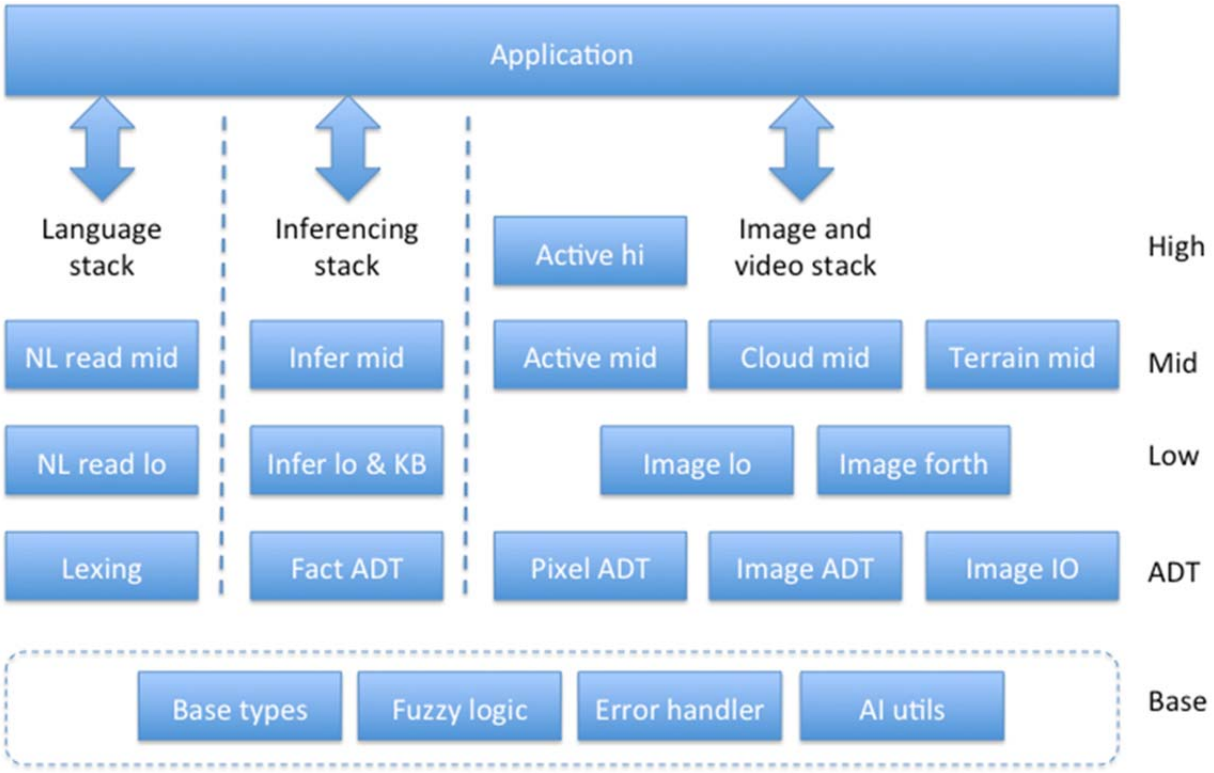


For resilience, the architecture is run cross-site duplicated and cross-linked, providing a hot backup. AutoPilot both schedules acquisition and analytic builds, and cross checks on the system health. AutoMonitor, not shown on the diagram, monitors health of the replicated system, and raises alerts if there are problems.

Powerful Multi-Technology Integrated Architectures

Multiple techniques to address complex questions

We use multiple AI (artificial intelligence) and data analytic techniques to solve the requisite problem. Here, for example, is an AI architecture designed to allow queries expressed as natural language emails, on spacecraft image data:



This architecture used natural language processing, image understanding, and inferencing with (as it happens) a partially self-deducing knowledge base. (In this case, the incoming email caused the system to induct the most appropriate knowledge base to use to then answer the question.) It was able to operate in multiple languages.

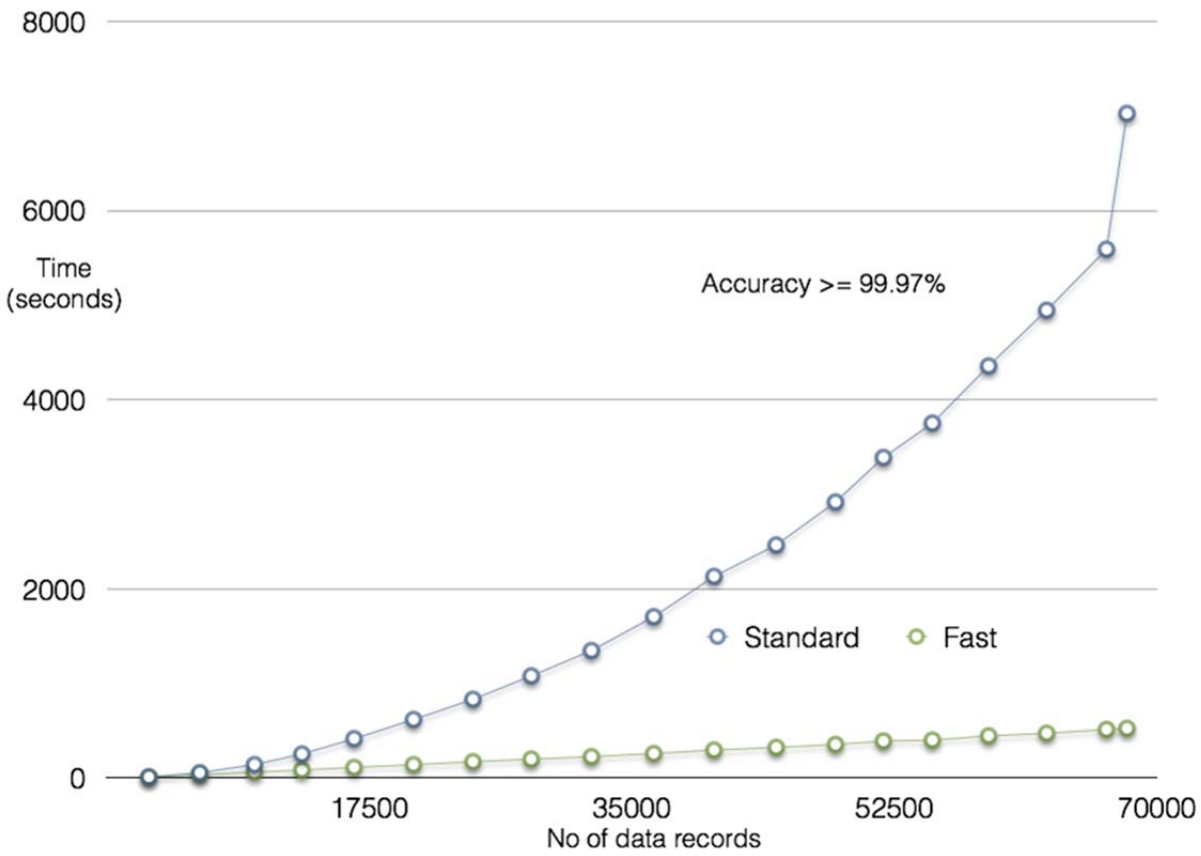
Order (N) Processing

We answer hard questions despite the growth in processing time implied by big data

It is an unfortunate property of the real world that the most useful questions to address are often the hardest, and so it is with data analytics. For example, it is frequently useful to know which of many items are like which others: perhaps, which financial transactions are so similar to other transactions that they indicate a potential fraud. For this type of question, there is really no avoiding the 'all-against-all' query.

Generally, data analytic houses will avoid even mentioning this type of query, since they *cannot do it on big data*. The reason is that it runs in *Order*(N^2): the time it takes is proportional to the *square* of the number of data items. However fast your processing, N^2 can always grow to large and with big data, it does.

We developed a smart heuristic which allows questions to be tackled in Order(N).
As a heuristic, it necessarily does not visit all pairs, so it can miss viable answers. The question is, how often does that happen? We conducted some exhaustive tests and measured the accuracy as never dropping below 99.97%, which of course is considerably greater than any similarity metric that might be employed anyway. And the alternative is – to simply have no answer. The graph shows the growth in execution time.



Using Primary Key Analytics

Our analytics at your service

You can use our analytics in several ways, for:

- Social media analytics, by taking our Illuminate service
- Fraud discovery, by taking our Distil service
- Custom analytics, including investigations and trial support, by engaging us on a consultancy basis
- Integrating our analytics into your systems and services, by discussing integration and licensing
- Using our skills to enhance your developments, by engaging our programming and design consultancy services

Primary Key Associates Limited

Data Analytics

About us

Since 2010 Primary Key Associates has delivered world-class consultancy, developing and exploiting cutting-edge analytics, artificial intelligence and digital investigation technologies.

Answer-Focused Technologies

We re-invest much of our revenue to develop cutting-edge data analytics, investigation and artificial intelligence technology to address the business problems we see our clients facing. Our IPR portfolio includes:

- Primary Key **Scenario Analytics** – a fast, modern and flexible data analytics technology to find, explain and illustrate the 'unknown knowns' in your large datasets.
- Primary Key **Insight Engine** – Scenario Analytics combined with powerful entity and relationship extraction, applied to digital forensics that reveals previously unknown connections in evidence.
- Primary Key **Illuminate** – a social media analysis and intelligence technology and service to find and analyse open source data to address particular business problems.
- Primary Key **Incipients** – predictive analytics that identify what is going to trend.
- Primary Key **Distil** – a technology toolkit we deploy on client sites to both find frauds and identify the business practices that make you vulnerable to fraud.
- **Fast analytic heuristics** to overcome the bottleneck of solving intrinsically non-parallel, yet vital, problems.

Expert consultancy and services

Our team are experienced professionals in IT and related areas:

- As experts in enterprise and security architectures (including SABSA), information security and cryptography we help design, validate, and test systems which are secure both physically and in cyber-space.
- We build software systems from the smallest (in C and assembler coded microcontrollers) through the mobile (Java or Swift coded Android or IoS apps), to the ubiquitous (Python or C+ on PCs, web technologies and virtualised servers) and even the esoteric (real-time and spacecraft).
- We understand data (SQL and graph databases) and artificial intelligence (computational linguistics, image analysis and machine learning) and have fast algorithms for timely and secure analysis of big datasets and evidential analysis of social media.
- We provide cyber threat intelligence and competitive intelligence.
- We support civil and criminal digital investigations from open source research through digital forensic analysis and providing expert witness reports and testimony.
- We undertake both technical and business programme and project management.

e: feedback@primarykey.co.uk
w: www.primarykey.co.uk
t: +44 1403 599900

@pkaluk