Big Data Integration: Challenges, Tasks and Tools

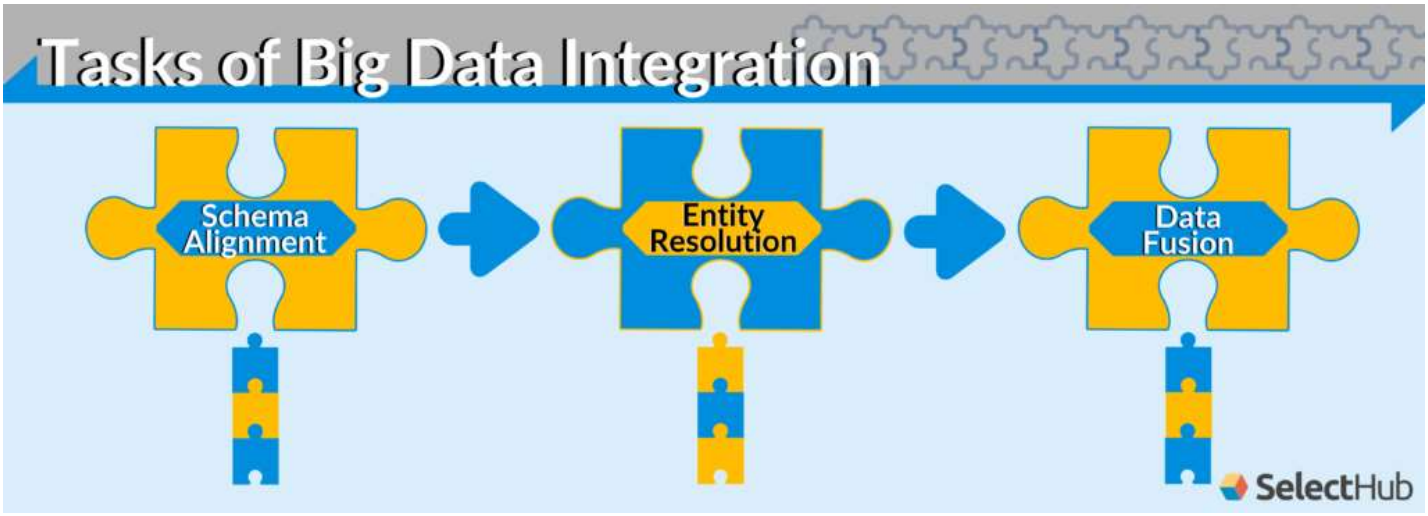
Richard Allen (https://www.selecthub.com/author/richard-allen/)

Big Data Analytics (https://www.selecthub.com/category/big-data-analytics/)

No comments (https://www.selecthub.com/big-data-analytics/big-data-integration/#respond)

So, you want to add big data tools (/big-data-analytics-tools/) to your business. And why wouldn't you? Big data analytics gives you a competitive edge, helps you optimize your operations and gives you a broader overview of your company. However, it's not as simple as snapping your fingers and telling your staff to implement BDA. Big data integration is a complex process with high rewards.

Get our Big Data Requirements Template (https://pmo.selecthub.com/big-data-requirements-onsite/)



(/#email) (/#facebook) (/#linkedin) (/#twitter) (/#print) (https://www.addtoany.com/share#url=

integration%2F&title=Big%20Data%20Integration%

It's not as simple as compiling all of an organization's structured operational data in a warehouse (/business-intelligence/data-warehouse-requirements-gathering/). It requires extracting data from a variety of sources, structured, unstructured or semi-structured, making it all compatible with each other, and then storing that data in a warehouse or lake where it can be accessed later.

If traditional data integration is a glass of water, then big data integration is a smoothie. Let's explain.

If you're thirsty, you can just take a glass, stick it under the faucet, turn a knob and bang, you're hydrated. Or, you can make a smoothie. You throw some yogurt, milk and ice cubes into a blender. Then you want to add some fruit. But you can't just throw in a full banana or some strawberries; you have to peel the former and cut the stems off the latter and maybe throw out a couple that went bad. Then you throw it all into a machine, let it do its thing, and voila, you've got a homogeneous liquid that nourishes not just your thirst, but gets you some essential vitamins and extra perks that your glass of water didn't have.

Sure, it took a lot more effort than just stepping up to the sink, but you got more for it. Such is the world of big data. Integrating your business's internal datasets with industry data can be make-or-break for some establishments. To make that data usable, coherent integration processes are a necessity.

In this article, we'll explore everything you need to know for getting your nourishing big data insights. We'll discuss the process of merging data, the challenges of scaling those efforts up to the level of big data, the questions you need to ask before integrating and tools for setting you on your way to enterprise analytics health.

Challenges

As is the case with most discussions in life, integrating big data often boils down to an internal debate between tangible resources vs. monetary cost. Many of the challenges that are presented in the big data process can be resolved by simply outsourcing the workload to a product or service. Some of the major challenges of integrating big data are:

- Finding skilled and capable big data engineers and analysts to develop workflows and draw actionable conclusions from the process.
- Ensuring the accuracy, quality and security of the data.
- Upscaling data-processing efforts.
- Synchronizing all data sources.
- Storing data effectively and efficiently.

There are four distinguishing characteristics of big data that separates it from "small" data: Volume, variety, velocity and veracity. Each of the Four V's (/business-analytics/crash-course-big-data/) present unique challenges of data integration.

Volume

Coordinating large amounts of data on its own is a challenge. To use big data, companies must dedicate extensive resources to data harvesting, processing and storing, either physically or financially. If your business doesn't have an extensive computing network, services like Hadoop provide outsourced processing. Recognized as one of the cheapest options for big data, individual nodes can still cost $4,000 (https://analyticstraining.com/can-big-data-solutions-using-hadoop-save-you-big-bucks/).

It starts to add up quickly, especially if your business is constantly streaming data and utilizing real-time metrics. Other than the cost, the logistics of dealing with all of that data can be a daunting task.

Variety

Perhaps the most significant component and consequently biggest challenge of big data integration is working with a variety of data.

While having lots of data is the superficial definition of big data, the true value comes from complex, deep datasets. Multidimensional data allows for deeper insight discovery than surface-level analysis of larger single-dimension sets.

Using more sources from individual silos, not bigger sources, an idea MIT professor Michael Stonebraker called the "Long Tail" of big data (https://sloanreview.mit.edu/article/variety-not-volume-is-driving-big-data-initiatives/), is the most essential component.

But making thousands of unique data sets, with different or no schemas, work together requires advanced analytics resources, capabilities and sophisticated knowledge of how to use them.

Velocity

If it takes weeks to process and produce insights on big data, odds are by the time all the work is done, the new knowledge gained from it is obsolete by the time it's in your hand.

More and more companies are relying on real-time analytics. Even those that don't need up-to-the-minute info still don't want to wait weeks or months to take action. In tandem with volume and variety, velocity becomes a challenge for integration.

When working with complex, large datasets, it's most likely impossible to apply a uniform analyzing process to it all. Because some individualizing is required, the task slows significantly. Big data integration tools like Alteryx (/big-data-analytics-tools/alteryx/) and Essbase (/big-data-analytics-tools/essbase/) allow for load balancing and distributed data processing, enabling different components of the set to be analyzed at the same time and increase speeds. But, again, that involves paying more money.

Veracity

According to a survey by Forrester Consulting (https://www.forrester.com/Global+Business+Technographics+Data+And+Analytics+Survey+2019/-/E-sus5171) in 2019, only 38% of business executives were confident in their worker's customer insights, and 34% were confident in business operations insights. That's because validating accuracy and relevance is a huge challenge in analytics, especially in big data.

Tasks

Integrating data follows a three-step process: schema alignment, record linkage then data fusion. It's a broad generalization of a sophisticated process. Still, it boils down to standardization of data organization between each source, finding points across sets that refer to the same entity and merging the data. It can then be stored in a warehouse for analysis. There are other subtasks (https://docs.informatica.com/integration-cloud/cloud-data-integration/current-version/tasks/data-integration-tasks.html) involved in integration, such as mass ingestion and replication, but these all fall into the major three steps.

The process gets modified with big data, complicating each step. Most notably, the inclusion of unstructured data adds a whole new step to the process, in that it doesn't follow a schema. The organization of the data needs to be developed from the ground up.

Giving Structure To That Which Has None

Organizing unstructured data is perhaps the most significant barrier to entry for big data analytics. The process of doing so varies from one data format to another.

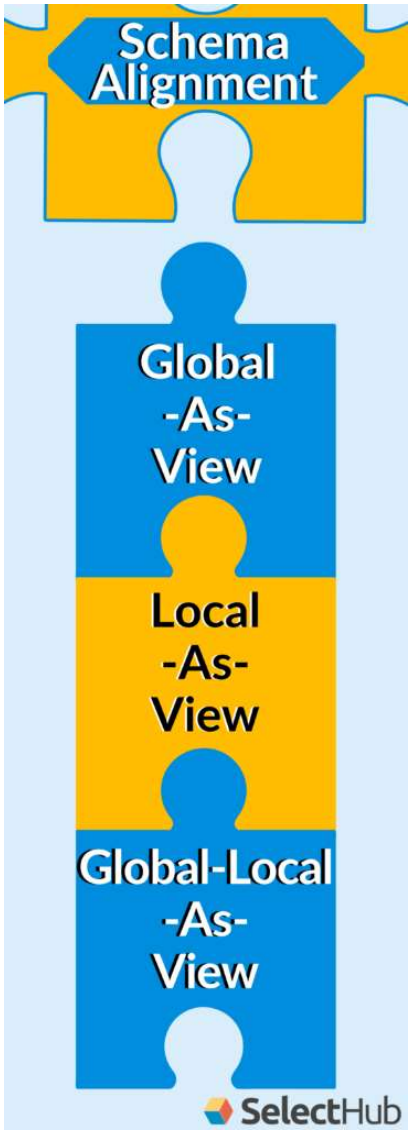
Text sources can use natural language processing to parse out individual words and phrases and, with the help of human guidance, assign semantics and organization to that data. For images and videos, it follows a similar process, using optical character and object recognition to define data in terms that fit into a schema.

Once the data is organized, it can start to be grouped and cleansed.

Schema Alignment

The first task of the data integration process is to uniformize the schemas of all datasets. This step includes three substeps: mediated schema creation, attribute matching and schema matching.

Mediated schemas are a uniform structure given to all data sources. It acts as a template to follow in the next two steps. These provide a consistent architecture for actual storing and subsequent analyzing, enabling universal functions on the complete set of data.



(https://www.addtoany.com/share#url=

integration%2F&title=Big%20Data%20Integration%

Once the mediated schema is created, attribute matching then takes place. Following the mediated schema, the source datasets are, as the name implies, reorganized to match data points to corresponding schema dimensions. It should be noted that this is often a one-to-one translation, but could also result in individual data points holding characteristics that match several attributes of the mediated schema. For example, the first name "John" from the source could be inserted into the mediated schema's "name1" dimension as well as its "fullname" dimension to compose "John Doe" with the last name.

Lastly, schema mapping is developed. It simply specifies the link between the mediated schema and the original data source. There are three types of maps: global-as-view, local-as-view and global-local-as-view (http://article.nadiapub.com/IJAST/vol120/3.pdf):

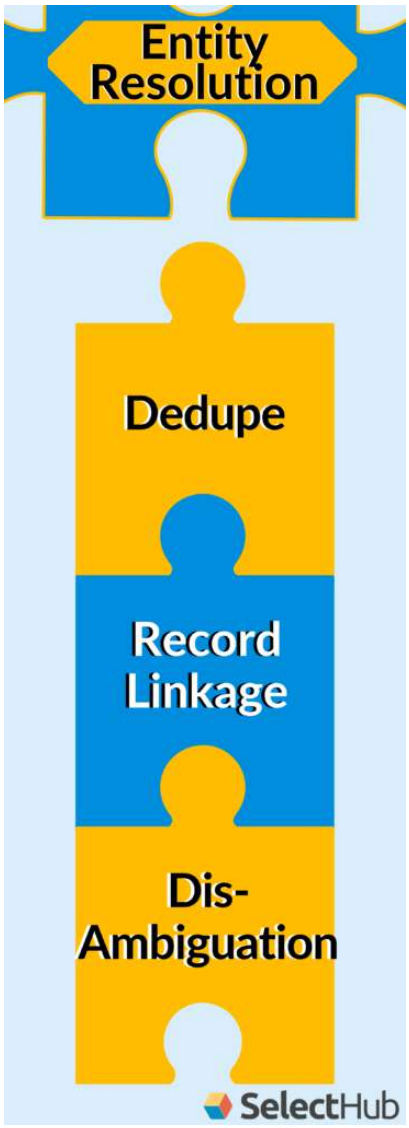
- **GAV:** Specifies how to find data in the mediated schema via the original source.
- **LAV:** Specifies how to find data in the original source via the mediated schema.
- **GLAV (also known as Both-As-View or BAV):** Allows two-way querying between the mediated schema and original source and vice versa.

LAV allows for easier adding of additional sources, while GAV provides more intuitive, quicker querying.

Schema alignment addresses challenges in the variety and velocity dimensions by uniformly organizing all datasets into one schema, which can be acted on by single processing functions queries.

Entity Resolution

Entity Resolution (https://www.districtdatalabs.com/basics-of-entity-resolution) is a data cleansing process. It involves semantically aligning pieces of data that relate to the same entity, omitting irrelevant entities, and disambiguation of noise. Essentially, it's optimizing the new uniform dataset formed in the schema alignment stage for accuracy and speed.



It also follows three steps:

- **Deduplication (https://www.dataversity.net/data-deduplication-can-help-reduce-cloud-costs/):** Removing exact copies of the same sets of data.
- **Record linkage (https://winpure.com/blog/what-is-record-linkage/):** Linking pieces of data that refer to a single entity. This can be matching state names when one source spells them out, another uses their abbreviations, and a third labels them by the number at which it joined the Union (rare, but it's probably happened at least once, right?).
- **Canonicalization/Disambiguation (https://searchdatamanagement.techtarget.com/definition/disambiguation):** Aligning ambiguous entities with clear ones to give semantics to noisy data. The abbreviation COL could refer to the country Colombia, the state Colorado, the military rank colonel, be short for the word color, or even be the plain old word. This step uses entity context and matching to specify the actual semantics of the data point.

This step, in "traditional" data integration, was simply referred to as record linkage. But the complexities of big data, including the introduction of poor-quality and repetitive datasets, have made data cleansing and optimizing more paramount than in historical integration efforts.

Entity resolution trims the volume of the data while informing veracity and confidence in data.

Data Fusion

Once all the data sources are organized and cleaned, it's time to mash it all together. Data fusion is the final step, where veracity and data quality gets hammered hard.

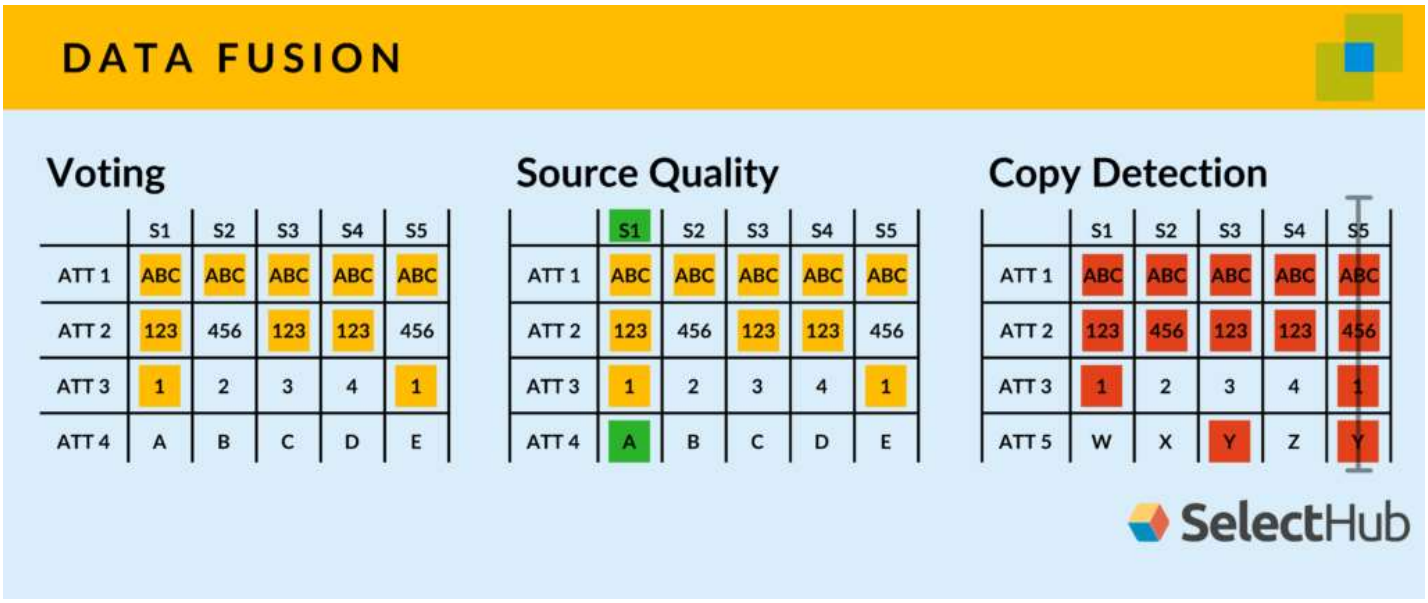
When merging the datasets, it's important to develop a hierarchy of trustworthiness and usefulness. Primary sources, or independent sources as they're referred to in the integration process, are typically more enriching and trustworthy than second-hand aggregators, or copiers in this context.

Data fusion is composed of three pieces:

- **Voting** compares values for an attribute across sources and finds the most common value for each.
- **Source quality** takes the information discovered in the voting process and determines which sources are the most "accurate" based on their production of the most common values. It then gives more weight to those sources.

- **Copy detection** identifies and removes copier sources. If one source produces only values that found in other sets, it is expendable. The accuracy of that set is irrelevant: whether its values are true or not, they are represented in other places.

The concept of data redundancy (https://www.talend.com/resources/what-is-data-redundancy/) is integral to the data fusion step. Using elements of one source that are repeated across others allows the accuracy of that dataset to be verified. If you can verify the repeated elements in that set, your confidence in the unverifiable data points increases. Before all that redundant data is cleared out for velocity and volume purposes, it can inform your nonredundant data and increase veracity.



Get our Big Data Requirements Template (https://pmo.selecthub.com/big-data-requirements-onsite/)

Approaches

For a long time, data integration with synonymous with extract, transform, load (https://www.sas.com/en_us/insights/data-management/what-is-etl.html). But the expansion of the data landscape has resulted in an expansion of methods, as well.

ETL is the general workflow of prepping data for analysis, whether it be big or small, integrated or siloed. Because it is a fairly generic term, it is scalable to big data.

Older varieties include a simple export and import approach and point-to-point integrations (https://www.informit.com/articles/article.aspx?p=28713&seqNum=2), both of which fell out of fashion because of their lack of scalability.

The popular, new kid on the block is data virtualization (https://www.datamation.com/big-data/what-is-data-virtualization.html). The big reason for its rise is its ability to query data and manipulate without having to directly pathway through to the original source. Data can be instanced in a virtual layer that can extend across applications and even devices, which allows load balancing and real-time analysis. In a nutshell, it lets you stream data efficiently and without sophisticated knowledge of its origin.

Because of its zero replication characteristic, no data is altered or duplicated from the source, increasing speeds and preserving the integrity of the source.

Questions to Ask

So what's it going to take to get big data integrated and working for your business? If you're looking to dive in, there are some questions you need to consider. Here's a short list to get started:

- What sources do I need to support?
- How much data do I need?
- Do I need to stream real-time data?
- What format do I need?
- How much do I want to pay?
- What insights do I want from the data?
- Is my reason for wanting to integrate big data feasible?
- Is it secure?
- Is it scalable to new emergent environments and new sources?

You need to consider not just your current data needs, but those in the future, as well. Just because your data needs are low now doesn't mean they'll stay there. You could discover one insight on a customer persona that warrants a whole new investigation on a market you had never considered before.

It's also worthwhile to invest time in finding the perfect vendor for your business. Just because an option is more expensive and comes with a more extensive suite and integrated products, doesn't mean it fits your needs best. Open-source tools can reduce costs and provide a lot of the same connectivity to data as top-shelf commercial options.

These ideas extend to which databases you'll want to use as well. Not all databases are compatible with all applications. Relational databases like MySQL (/miscellaneous-software/mysql/) and NoSQL databases like MongoDB (/big-data-platform-software/mongodb/) will require different connectors and API to be analyzed.

Get our Big Data Requirements Template (https://pmo.selecthub.com/big-data-requirements-onsite/)

Next Steps

In this article, we discussed the challenges, tasks and details of big data integration. We went in-depth on potential approaches to big data integration, and how big data's four V's distinguish big data from traditional, "small" integration.

If you're looking to take the next step in adding big data to your enterprise, our experts at SelectHub are ready to help. Our requirements template (https://pmo.selecthub.com/big-data-requirements-onsite/) lets you start your search with a focus on what you need, and our requirements and features outline (/big-data-analytics/big-data-analytics-requirements/) can let you know what to look for in a product. If all this big data jargon is still making you scratch your head, our "What is Big Data? (/category/big-data-analytics/)" and crash course (/business-analytics/crash-course-big-data/) articles are good spots to start building up your skillset. And if you're ready to look at a comprehensive comparison of big data integration tools (), we've got you covered there, too.

What further questions do you have about big data integration? Did we miss anything? What kinds of data have you put together for your business? What challenges did you have in the integration process? Let us know below in the comments.

(/#email) (/#facebook) (/#linkedin) (/#twitter) (/#print) (https://www.addtoany.com/share#url=https analytics%2Fbig-data-integration%2F&title=Big%20Data%20Integration%

Leave a Reply

Your email address will not be published. Required fields are marked *

Your message

Your name *

Your email *

Save my name, email, and website in this browser for the next time I comment.

Post Comment

Search

Compare the Top Big Data Analytics Tools

Pricing, Ratings, and Reviews for each Vendor. PLUS... Access to our online selection platform for free.

(https://pmo.selecthub.com/request-custom-leaderboard/?category_slug=big-data-analytics-tools&product_slug&slug=big-data-analytics-tools&category=Big%20Data%20Analytics%20Tools&scorecard_id=266)

Requirements Template for Big Data Analytics Tools

Jump-start your selection project with a free, pre-built, customizable Big Data Analytics Tools requirements template.

(https://pmo.selecthub.com/big-data-requirements-onsite/)

- Applicant Tracking Systems (/applicant-tracking-systems/)
- Big Data Analytics (/big-data-analytics-tools/)
- Business Analytics (BA) (/business-analytics-tools/)
- Business Intelligence (BI) (/business-intelligence-tools/)

Business Phone (/business-phone-systems/)

Call Center (/call-center-software/)

Compensation Management (/compensation-management-software/)

Construction Bidding (/construction-bidding-software/)

Construction ERP (/construction-erp-software/)

Construction Estimating (/construction-estimating-software/)

Construction Management (/construction-management-software/)

Construction Scheduling (/construction-scheduling-software/)

CPQ (/cpq-software/)

CMMS (/cmms-software/)

CMS (/cms-software/)

CRM (/crm-software/)

Customer Experience (/customer-experience-software/)

Dental (/dental-software/)

Distribution (/distribution-software/)

Dispatch (/dispatch-software/)

Ecommerce (/ecommerce-platforms/)

EDI (/edi-software/)

EHR (/ehr-software/)

EMR (/emr-software/)

Employee Scheduling (/employee-scheduling-software/)

Embedded Analytics (/embedded-analytics-tools/)

Enterprise Accounting (/accounting-software/)

EAM (/eam-software/)

Endpoint Security (/endpoint-security-software/)

Enterprise Reporting (/enterprise-reporting-system/)

ERP (/erp-software/)

ETL (/etl-tools/)

Facility Management (/facility-management-software/)

Fundraising (/fundraising-software/)

Field Service Management (FSM) (/field-service-software/)

Fleet Management (/fleet-management-software/)

HR Management (/hr-management-software/)

Help Desk (/help-desk-software/)

Home Health (/home-health-software/)

Hotel Management (/hotel-management-software/)

Inventory Management (/inventory-management-software/)

Insurance (/insurance-software/)

Legal (/legal-software/)

Live Chat (/live-chat-software/)

LMS (/lms-software/)

Long Term Care (/long-term-care-software/)

Manufacturing (/manufacturing-software/)

Marketing Automation (/marketing-automation-software/)

Medical Billing (/medical-billing-software/)

Medical (/medical-software/)

Mental Health (/mental-health-software/)

Medical Practice Management (/medical-practice-management-software/)

MES (/mes-software/)

Patient Scheduling (/patient-scheduling-software/)

Payroll Systems (/payroll-software/)

Performance Management (/performance-management-software/)

POS (/pos-software/)

Procurement (/procurement-software/)

PLM (/plm-software/)

Property Management (/property-management-software/)

Project Management (/project-management-software/)

PPM (/ppm-software/)

PSA (/psa-software/)

Recruitment & Staffing (/recruiting-software/)

Risk Management (/risk-management-software/)

Sales Force Automation (/sales-force-automation-software/)

Supply Chain Management (/supply-chain-management-software/)

Takeoff (/takeoff-software/)

Talent Management (/talent-management-system/)

Telemedicine (/telemedicine-software/)

TMS (/tms-software/)

Time and Attendance (/time-and-attendance-software/)

Warehouse Management (/warehouse-management-software/)

Workforce Management (/workforce-management-software/)

Show All Categories (/categories/)

Why SelectHub (/why/)

Create a Project (https://app.selecthub.com/projects/new)

Browse Products (/categories/)

Dashboard (https://app.selecthub.com/dashboard)

Managed Selection Services (https://www.selecthub.com/managed-selection-services/)

Claim Your Product Listing (https://pmo.selecthub.com/claim-your-product/)

For Vendors (/seller/)

Thought Leader Program (/thought-leaders/)

Awards Program (/awards/)

Careers (https://www.selecthub.com/careers/)

About Us (https://www.selecthub.com/about/)

Privacy Policy (/policies/)