Q: When I think of Watson BigInsights, I think of support for querying and analyzing Hadoop big data sets. What is the use case here for Cognitive Capture e.g. using Big Data to build a knowledge base for Content Classification?

A: We are using text analytics and content classification within Datacap to process documents. It is being used to find data and classify documents, pages, and sections of text. We do not use BigInsights -- we use technology that is contained in BigInsights -- i.e. Text Analytics entity extraction.

Q: Are annotators and dictionaries provided via the Watson BigInsights runtime platform?

A. We are using the BigInsight Text Analytics runtime for text extraction.

Q: Will the Insight Edition add-on require any type of indexing server as well as a separate Cognitive processing server?

A: No.

Q: Has IBM Content Classification been extended in any way to support the Datacap Insight Edition e.g. knowledge base integration with Watson BigInsights, pre-defined classification rules for Cognitive capture, etc.?

A: No changes to IBM Content Classification. Datacap is using more features of CC like decision plans.

Q: How do we configure Taskmaster Web 8.1 outside the firewall?

A: When the web server is configured outside the corporate firewall observe the following for configuring Taskmaster Web:

- Any database accesses performed by Rules (e.g. validation rules on Verify screens) access the database directly. So the ODBC driver must exist on the IIS machine, the IIS application pool account (or the connection string) must have permissions for the database, and the firewall must permit TMWeb to access the database.
- All other database accesses and file accesses performed by TMWeb are done via TM Server and so the firewall ports open for TMS (typically 2042 and 2043) are sufficient.

Q: Since Datacap supports double-byte characters, why are not all Asian and similar languages supported?

A: Having support for double-byte character sets does not necessarily mean that all multi-byte, especially other Asian languages, can be automatically supported. The issue with some Asian languages is that they have "compound" characters, which require representation using multiple double-byte characters. So, for example, there may be characters which display as a

single item but actually consist of up to three double-byte internal characters -- for a total space of six bytes in the internal representation. This capability must be evaluated for each specific language. Without additional coding, many Datacap features --- such as the actions used to validate string lengths and formatting -- will incorrectly interpret these characters as three separate characters rather than one and fail to give the correct results.

Q: My customer wants a statement on the expected success rate of the Datacap OCR engines, and font and size recommendations for best results. Can IBM provide that?

A: IBM does not provide accuracy statistics and typically none of the vendors do. The customer is going to need to do testing on their own documents or have their Business Partner do this. If they want IBM to do this sort of testing it could be done by ECM Services as a billable engagement.

You should not be providing specific accuracy ranges or suggestions of what could be possible. There are too many factors outside of our control that depend on their specific environment and documents. Besides the font and size, the results can depend on the type of scanner, how well they do the scanning, fax vs scanning, the quality of the print on the page, paper color and background and condition of the pages, etc. So simply looking at the font size is not sufficient.

Giving a specific number will cause a problem if they have a different result, even if you qualify your statement with additional conditions, you will need to explain and defend your position and spend a lot of effort determining why they have particular results that vary from your statement. This is a very unproductive situation and needs to be avoided by not supplying accuracy numbers.

You need to work with the Datacap trained CTPs on the partner support group to help. They have worked with the product and would be your first line for assistance. Also, please use the Capture and Imaging community forum for questions whenever possible.

Q: What are the benefits of the Datacap Enterprise license when it comes to Rulerunner running in a multi-threading environment?

A: If the customer is going to use Rulerunner Enterprise, it usually costs much less than the old PVU-based license. If you do not use Rulerunner Enterprise, the system will only use one CPU core for running OCR/ICR, barcode recognition, and other functions. The system only processes one batch at a time on one core. With a 4 core server, running Rulerunner enterprise, you could run 4 batches simultaneously, where each batch can use a CPU core. This would increase the throughput by 4 times.

Q: Does DotEdit 8.1 support field-level validation out of the box?

A: No Datacap DotEdit does not provide this capability out of the box. This type of validation would need to be implemented as custom code in the verify panel itself upon exit for

example. In the out of the box software, validations are programed to run when submitting the current form or tabbing out of the last field on the form.

Q: How is Arabic supported in Datacap?

A: Via integration with NovoDynamics NovoVerus. See this FAQ and presentation for more information.

UPDATE: Arabic is also now supported out of the box using the (ABBYY) OCR/A engine.

Q: What's Datacap release history?

A: See Datacap Release History.

Q: Please, can you describe Datacap's tasks affinity with transactional APIs and Rulerunner servers?

A: When you do transactional capture services you do not use a Rulerunner server: the processing is done on the WTM server directly, and you load balance WTM servers using IP load balancing. For batch processing, using traditional Datacap, you configure tasks to Rulerunner servers but a task can be configured to to run on many servers. The Rulerunner service polls the Datacap (f.k.a Taskmaster) server to fetch work from the queue. Each Rulerunner "thread" is a process, so if you configure 8 threads, you will process 8 tasks simultaneously.

Q: What languages and check standards are supported for Check Processing?

A: US, UK, Canada, Brazil, France, Argentina, and India checks are supported today. The following additional countries/languages can be qualified: Australia, Chile, Columbia, Dominican Republic, Ecuador, Italy, Malaysia, Papa New Guinea (PNG), Portugal, and Puerto Rico.

Q: What's included in Datacap on Cloud and what are the terms of service?

A: Information is provided at two levels: specific product offering and the general terms that apply to all products. Start with the information in the section Datacap on Cloud on ZACS, for a high-level description. Then have a look at the Service Description for Datacap and the General Terms of Use for all cloud offerings.

Q: In Datacap on Cloud, how is the outbound bandwidth counted?

A: The outbound bandwidth is a monthly measure. Backups and data replication that we do are not counted.

Q: In Datacap on Cloud, is there a limit on the inbound bandwidth?

A: No inbound limit.

Q: In Datacap on Cloud, how frequently is IBM applying patches and updates?

A: For patches: we scan our environments for patch requirements and then apply based on level of criticality. In general, our objective is to apply required patches under a month, with emergency / high priority patches being applied sooner.

For Version Upgrades: we periodically review the environments and determine when it would be advisable to apply a version upgrade. Since these are single tenant environments we do have the ability to coordinate the scheduling of such items with our clients for testing and avoidance of peak periods.

Q: What are the recognition engines included in Datacap (base and optional) and how are they licensed?

A: Datacap Base includes the following recognition engines:

- **ABBYY** (OCR-A) for recognition of machine prints
- **Nuance** (OCR-S) for recognition of machine prints
- **OpenText RecoStar** (ICR-C) for recognition of hand prints
- **Tesseract** (Datacap Mobile) for recognition of machine prints

There is no limitation based in page counts for these engines.

Optionally, customers can also purchase and install separately the **NovoVerus** OCR engine from **NovoDynamics**, as an alternative OCR engine for Arabic. NovoVerus is licensed on a per machine and page count basis, plus a 20% annual maintenance and support fee. Contact NovoDynamics for details.

See list of supported languages here.

Optionally, customers can also purchase the following separately charged add-on components based on the **Parascript** engine:

- Datacap Check Processing Check Packs
- Datacap Cursive Recognition
- Datacap Signature Validation

Check processing is sold by packs of 10,000 per year. There is no limitation based on unit counts for Datacap Cursive Recognition and Signature Validation; they are sold based on PVU.

See the Product Announcement for details.

Also see Licensing Terms here.

Q: What is the definition of the recognition confidence levels in Datacap?

A: I am afraid we do not have a precise definition for each confidence level. Datacap normalizes recognition confidence levels assigned by OCR engines between 1 and 10, 10 being the highest. These values are just an estimate of the probability that a given glyph shown in the image is actually the character it is supposed to represent. The lower the quality (low resolution, low contrast, high background, smears, etc.) of the image, the lower the confidence level. These are relative levels that can be compared to a threshold. As you know, they are used in the Verify user Interfaces of the Datacap clients and in some actions, especially the ChkConfidence action. In the Verify user interface, by default, any value lower than 10 will cause the associated character and field to be flagged for review, which in effect is biased toward caution, as you want to make sure as few errors as possible go undetected. In most cases, when the quality of the scanned document is good, OCR engines will assign the max confidence level. But since it is a probability value computed by the OCR engine based on its internal recognition logic, there is the possibility of false positives, when the engine assigns a high confidence level to a wrongly recognized character (as in substitution cases "0" for "o" "1" for "I" or "l", etc.), and false negatives (more often), when assigning a lower confidence level to a correctly recognized character. So, to relax the default bias toward caution, you can tweak the "acceptable" confidence level threshold value for each field type, once you have gained experience with the quality of the extracted data, after a while in operations, so that false negatives above the threshold are no longer flagged for review, hence reducing the burden of verification.

See standard documentation
https://www.ibm.com/support/knowledgecenter/SSZRWV_9.1.4/com.ibm.dc.develop.doc/dcdev327.htm

Also, refer to page 48 and other places in the Datacap Redbook for more on confidence levels and factors affecting quality of the recognition process:
https://www.ibm.com/support/knowledgecenter/SSZRWV_9.1.4/com.ibm.dc.develop.doc/dcdev327.htm

Q: What is the difference between 5 and 6 confidence levels, logically?

A: Well, 5 is worse than 6 :-). By default, both are considered low confidence since less than 10. Now, if you set the threshold at 6 for the field, then characters with a confidence level of 6 and above will show as "passing", and those with 5 and less will be flagged for review. So before you do this, you need to make sure, through practical experience with your documents, that you are letting through only false negatives.

Q: How can customers trust the results of each confidence level?

A: As explained earlier, by running production documents through the system with the default, systematically checking the recognition results for a while on a representative set, and then if necessary by gradually lowering the threshold to reduce the amount of manual handling without affecting the error rate. By then you will likely have identified the typical types of errors and applied corrective actions in your application.

Q: What is the difference between ABBYY and Nuance in the logic of calculating Confidence levels?

A: We can only comment from a Datacap perspective, as the processing logic and confidence calculation method of each engine is proprietary to each vendor. As indicated above, each engine's confidence level scheme is normalized in Datacap. We typically get very similar results for both, with some marginal differences in speed vs. accuracy. A given set of documents, in a given language, will typically yield very similar OCR results and confidence levels for both.

Q: I would like to know the logic of 10 and how confidence of 10 if possible, exact match logic comparing each OCR engine character template etc?

A: To my knowledge most OCR engines, commercial or Open Source in the market, have this mechanism of confidence level. My understanding is that each OCR engine applies various techniques depending on the many characteristics exhibited by an image, including quality of the image, character and font attributes, page layout, language, etc. and the OCR processing logic that has been applied to the image. For example, OCR engines expect images within a certain resolution range, typically with best results between 200 and 300 DPIs. This is to help it analyze a given image and segment/identify the zones of interest in it, such as zones with lines of text, font types and sizes, zones with pictures, zones with tables, etc. so as to apply the appropriate character recognition process in subsequent steps. For example, knowing the DPI resolution (dots per inch) of a given image, an OCR engine can compute the size of certain artifacts in a text zone and determine whether it is dealing with background noise (speck) or part of a character, as in a doted "i", in a small sized font which, guess what, is also determined based on the DPI...

Then OCR engines isolate individual characters in text zones and compare them to a collection of character bitmap templates (in various fonts and sizes). As a result, the OCR engine assigns each character a recognition confidence level based on how well it correlates with the template/matrix variants and also how well ambiguities can be resolved using dictionaries or lexicons. So, yes, OCR engines will assign their max confidence level, translating to 10 in Datacap, when there is a perfect match with one of the character variants, for the right font type and size, and when the character, together with the ones surrounding it in a word, is found in the OCR language dictionary. Note that there is a lot of interactions with the processing context and many iterations to arrive at the final confidence level output by the engine, and this is why OCR is so compute-intensive.

Q: If, by default, any value lower than 10 causes the associated character and field to be flagged for review, does that mean Datacap trusts 10 then?

Yes, Datacap itself does not have an alternative way of assessing confidence levels of its own, and basically trusts OCR engines when they say they got it right. But again, note that there is a caution bias built into the system, as more often than not, OCR engines return confidence levels lower than 10, even if they recognize the right characters (for example: because the DPI does not allow to determine an exact match for the font size, because there are a couple of specks that

make a "c" partially match a "e" bitmap, but "c" can still be found at that position in "practice" but not in "praetice", which does not exist, etc.).